Census 2000 Testing,
Experimentation,
and Evaluation Program

October 21, 2003

Synthesis of Results from the Administrative Records Experiment in 2000 (AREX 2000)

FINAL REPORT

This research paper reports the results of research and analysis undertaken by the U.S. Census Bureau. It is part of a broad program, the Census 2000 Testing, Experimentation, and Evaluation (TXE) Program, designed to assess Census 2000 and to inform 2010 Census planning. Findings from the Census 2000 TXE Program reports are integrated into topic reports that provide context and background for broader interpretation of results.

Dean H. Judson and Barry Bye
Planning, Research, and Evaluation Division

Intentionally Blank

ACKNOWLEDGMENTS

The Administrative Records Experiment 2000 was conducted by the staff of Administrative Records Research at the U.S. Census Bureau, led by Charlene Leggieri. Questions and comments regarding this document can be directed to Dean H. Judson at 301-763-2057.

Administrative Records Research Staff Members and Key Contributors to AREX 2000:

Bashir Ahmed
Mikhail Batkhan
Mark Bauder
Mike Berning
Harold Bobbitt
Barry Bye
Benita Dawson
Joseph Conklin
Kathy Conklin
Gary Chappell
Ralph Cook
Ann Daniele
Matt Falkenstein
Eleni Franklin
James Farber
Mark Gorsak
Harley Heimovitz
Fred Holloman
David Hilnbrand
Dave Hubble
Robert Jeffrey
Dean Judson
Norman Kaplan
Vickie Kee
Francina Kerr
Jeong Kim
Myoung Ouk Kim
Charlene Leggieri
John Long
John Lukasiewicz
Mark Moran
Daniella Mungo
Esther Miller
Tamany Mulder
Nancy Osbourn
Arona Pistiner
Ron Prevost
Dean Resnick
Pamela Ricks
Paul Riley
Douglas Sater
Doug Scheffler
Kevin A. Shaw
Kevin M. Shaw
Larry Sink
Diane Simmons
Amy Symens-Smith
Cotty Smith
Herbert Thompson
Deborah Wagner
Phyllis Walton
Signe Wetrogan
David Word
Mary Untch

and

Members of the AREX 2000 Implementation Group

Intentionally Blank

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Intentionally Blank

EXECUTIVE SUMMARY

The Administrative Records Experiment 2000 was an experiment in two areas of the country designed to gain information regarding the feasibility of conducting an administrative records census or the use of administrative records in support of conventional decennial census processes. The first experiment of its kind in the United States, the Administrative Records Experiment 2000 was part of the Census 2000 Testing, Experimentation and Evaluation Program. The focus of this program was to measure the effectiveness of new techniques for decennial census enumeration. There were four evaluations: Process, Outcomes, Household, and Request for Physical Address evaluations. The first three are summarized here.

Administrative Records Census definition and requirements

In the Administrative Records Experiment, an administrative record census was defined as a process that relies primarily, but not necessarily exclusively, on administrative records to produce the population count and content of the decennial census short form with a strong focus on apportionment and redistricting requirements. In addition to total population counts by state, the decennial census must provide counts of the voting age (18 and over) population by race and Hispanic origin for small geographic areas, currently in the form of census blocks.

Demographically, the Administrative Records Experiment provided date of birth, race, Hispanic origin, and sex. Geographically, the Administrative Records Experiment operated at the level of basic street address and corresponding Census block code. Unit numbers for multi-unit dwellings were used in certain address matching operations and one of the evaluations; but generally household and family composition were not captured. The design did assume the existence of a Master Address File and geographic coding capability similar to that available for the Census 2000.

The principal objectives of Administrative Records Experiment 2000 were twofold. The first objective was to develop and compare two methods for conducting an administrative records census, one that used only administrative records and a second that added some conventional support to the process in order to complete the enumeration. The second objective was to explore the potential use of administrative records data for some nonresponding or unclassified households that occur in a conventional census.

Administrative Records Experiment Top-down and Bottom-up methods

A two-phase process accomplished the Administrative Records Experiment 2000 enumeration. The first, or Top-down, phase involved the assembly of records from a number of national administrative record systems and unduplication of individuals within the combined systems. This was followed by computer geocoding of street addresses to the level of census block and two attempts to obtain and code physical addresses for those that would not geocode by computer. Finally, there was a selection of “best” demographic characteristics for each individual and "best" street address within the experimental sites.

The second phase of the Administrative Records Experiment 2000 design was an attempt to complete the administrative-records-only enumeration by the correction of errors in administrative records addresses through address verification (a coverage improvement analogue) and by adding persons missed in the administrative records (a nonresponse followup analogue). Considering the Top-down and Bottom-up processes as part of one overall design, the Administrative Records Experiment can be thought of as a prototype for a more or less conventional census with the initial mailout replaced by a Top-down administrative records enumeration.

Limitations

There were four principal limitations on the experiment.

- The administrative records source files were limited to those used in the creation of the Statistical Administrative Records System 1999, which relied primarily on files for tax year 1998 and other files extracted early in calendar year 1999. These files neither exhausted the national-level administrative records that might have been available for Administrative Records Experiment 2000 nor were they the timeliest with respect to April 1, 2000, Census Day for Census 2000.

- The number of experimental sites was small. Although it would not have been reasonable or realistic to attempt to mount this first Administrative Records Experiment in a representative sample of geographic areas large enough to make national estimates, additional sites would have provided more confidence that the results were not idiosyncratic to the sites selected.

- There was no experimental variation in key design parameters such as the clerical and field operations and the address selection algorithm. Without some factorial or fractional factorial structure, direct estimates of operational impacts of components, individually or in combination, were not possible.

- The measurement of race and Hispanic origin in administrative records at the national level is deficient. Attempts were made to improve the measurement through the use of certain statistical models, but the results were not entirely satisfactory.

The limitations in the Administrative Records Experiment were largely due to resource constraints and a short planning period for what was an extremely complex and novel undertaking.

Experimental sites

Two sites were selected that were believed to have a total of approximately one million housing units and a population of approximately two million persons. One site included Baltimore City and Baltimore County, Maryland. The other site included Douglas, El Paso and Jefferson Counties, Colorado. The sites provided a mix of population and housing characteristics needed to assess the difficulties that might arise in conducting an administrative records census.

Administrative Records Experiment Outcomes Evaluation

As expected, the Bottom-up coverage is much improved compared to the Top-down, and this is largely due to the completion of the Top-down enumeration by using census data for nonmatched addresses, which simulates a followup to the administrative records enumeration. Specifically, the Bottom-up coverage of children (81 percent - 94 percent across the test sites) is substantially better than the Top-down (72 percent - 83 percent). Coverage of children is a particular weakness for administrative records used in Administrative Records Experiment 2000.

Adults in the Bottom-up are more or less uniformly overcounted (102 percent - 104 percent). The overcount of adults most likely is due to unaccounted for deaths in the 12 months prior to Census Day, the lack of special populations operations in the Administrative Records Experiment (e.g., a group quarters enumeration), and failure to unduplicate persons after adding census data for nonmatched addresses. Of course, the latter means that there is some duplication of children as well.

Detailed enumeration results focused mainly on a comparison of the Bottom-up enumeration with the Census 2000. The analysis did not include group quarters and, due to limitations in the administrative records sources, persons could not be reported with “multi” or “other” race. The analysis progressed from large geographic areas to small geographic areas, beginning with the five test site counties and ending with census blocks within the sites. The evaluation incorporated a variety of methods to accomplish its objectives, including univariate and multivariate statistical analyses of the Administrative Records Experiment/Census 2000 differences, and spatial/ecological maps that examined the geographic distributions of key comparison measures. The outcomes evaluation tried to disentangle the influence of demographic change and Administrative Records Experiment processing, coverage and data quality issues, while presenting basic enumeration statistics.

At the county level, the Bottom-up process undercounted total population in all sites except Baltimore City. As with the total population, males and females were undercounted in all sites except Baltimore City, but the female undercounts were slightly greater than male undercounts. Age groups showed more variability with most groups undercounted. Generally the size of the undercounts increased with decreasing age, except for the 20-24 age group. These patterns did not appear to be site-specific. Overcounts for the oldest old and undercounts for the youngest persons suggest that much more timely birth and death information must be obtained. Also, the special enumeration requirements for populations such as college students, the military and persons in nursing homes must be incorporated into administrative records processes.

Administrative records are not currently a good source of data for race and Hispanic origin, and the models were not sufficient to correct their deficiencies. Blacks and Hispanics were undercounted when they were a large minority group and overcounted when they were not. American Indians and Alaskan natives were not well identified and the accuracy of Asian/Pacific Islander counts was uncertain.

Bottom-up tract-level total population results indicated a good correspondence between Administrative Records Experiment and the census. The population counts of 70 percent of tracts were within 5 percent points, and 95 percent of the tracts were within 25 percentage points, though a sizable number of tracts had moderate and large undercounts. At the block-level, population counts were the least accurate. For the total population 38 percent of blocks met the 5 percent criterion and about 85 percent of blocks met the 25 percent criterion.

A multivariate analysis of block differences showed that large undercounts were associated with such block characteristics as high population density, high rental rates, and large proportions of persons age 20-24. Large overcounts were associated with high vacancy rates, low population density, small proportions of persons under the age of 20 and large proportions of persons age 20-24 and age 65 and over (Heimovitz, 2002).

Administrative Records Experiment Household-level analysis

The general goal of the household-level analysis was to assess how well households formed from administrative records matched those from Census 2000 addresses. The evaluation focused, first, on the factors associated with Administrative Records Experiment and Census 2000 addresses that were (computer) linked. Then, demographic comparisons were made between households at linked addresses. There was a special focus on Census 2000 households that required a nonresponse followup and Census 2000 unclassified (imputed) households.

The evaluation used both descriptive analyses and logistic regression analysis to assess the coverage and accuracy of Administrative Records Experiment households. Descriptive analyses were performed for households in all five Administrative Records Experiment counties and for the Census 2000 nonresponse followup and imputed households in the test sites. A logistic regression model was developed to predict the probability of an accurate household match using address and Administrative Records Experiment processing characteristics as predictors. Addresses with a high probability of correct demographic match between occupants might be candidates for administrative records substitution in the case of nonresponse followup in a conventional census. In the following discussion the term "linked" is used to mean a matched address. The term "matched" is reserved for household demographic comparisons at linked addresses.

Administrative Records Experiment's coverage of the census nonresponse followup universe was not as good as its coverage of the overall universe. Administrative Records Experiment housing units were linked with 70.9 percent of the census nonresponse followup housing units, compared with 88.4 percent of the census responding housing units. For occupied nonresponse followup housing units, the coverage rate was 76.7 percent. The Administrative Records Experiment housing units were linked with 63.2 percent of households that were imputed to have people in them, and 34.7 percent of those imputed to be vacant.

The Administrative Records Experiment and the census counted the same number of people in the housing unit for 51.1 percent of the 889,638 linked households, and Administrative Records Experiment was within one of the census for 79.4 percent of the units. The 51.1 percent is effectively a ceiling on the percent of linked households that had exactly the same persons from Administrative Records Experiment and Census 2000. Although errors in address linkage would account for some of the mismatched households, the deficiencies in administrative records cited earlier in this report--missing children, lack of special population operations and the time gap between the administrative records extracts and Census Day--most likely account for the major part.

For linked nonresponse followup housing units, Administrative Records Experiment had the same numbers of persons for 37.0 percent of the units and was within one 69.3 percent of the time. Census 2000 nonresponse followup housing units were more susceptible to the Administrative Records Experiment deficiencies than responding units. In addition, enumeration errors in Census 2000 might have been higher for these units.

The regression analysis demonstrated a number of factors associated with greater probability of matched household demographics. These include: single unit address rather than multi-unit, household with only one or two members, all household occupants over the age of 65, at least one White occupant, no occupant with imputed race in the Administrative Records Experiment. The predictive power of the model was moderately strong. At a predicted probability of 0.5 or higher, the probability of a correct household match was about 72 percent. At a predicted probability of 0.8 or higher, the probability of a correct match increased to about 83 percent, but the proportion of addresses with predicted probability this high was only about 4 percent of all addresses. Evidently, the limitations in the data, particularly the administrative records cutoffs and poor race and Hispanic origin measurement, made household prediction quite difficult.

Implications for 2010 Planning

Substitution for 2010 nonresponse followup households should continue to be explored

Although the results of the household-level analysis were not definitive due to the limitations on Administrative Records Experiment 2000, they were sufficiently strong that research into the substitution of administrative records households for nonresponse followup or unclassified households in a conventional census should continue. For nonresponse followup households there is the potential for significant cost savings, and for unclassified households, the potential for greater accuracy than that provided by imputation.

The approach piloted in the Administrative Records Experiment 2000 should be tested as part of the 2004 Census Test using models developed from a linkage of Statistical Administrative Records System 2000 data to the Census 2000 files. The timing of the administrative records in the Statistical Administrative Records System 2000 would be much closer to Census Day than the Statistical Administrative Records System 1999 data used in the Administrative Records Experiment 2000, and much more like the data that could be acquired for 2010.

Other 2010 impacts should be considered

There are other aspects of 2010 Census development in which administrative records might play a role. These include Master Address File improvements, development and testing of unduplication methods for 2010, subnational Demographic Analysis, and coverage measurement research.

2010 data acquisition and research agenda

Arrangements should be made to acquire administrative records on a timelier basis and to obtain some data sets that might fill some of the administrative records coverage gaps.

A research agenda for 2010 would include:

- Additional evaluation of the impact of clerical and field operations in Administrative Records Experiment 2000.
- Person unduplication in the Administrative Records Experiment Bottom-up process.
- Repeating Administrative Records Experiment 2000 with Statistical Administrative Records System 2000 data.
- Repeating the Household-level analysis using Statistical Administrative Records System 2000 data.
- Analysis of administrative records coverage gaps, in particular gaps related to persons in group quarters.
- Master Address File improvements using administrative records.
- Improving address linkage techniques.
- Enhancing Numident race and Hispanic origin data using Census 2000.
- Contributing to subnational Demographic Analysis.

Implications for other Census Bureau programs

The research that went into the development of the Statistical Administrative Records System and Administrative Records Experiment 2000 has had significant payoffs in Census programs other than the decennial census, and the development of new uses for administrative records should continue to benefit non-decennial programs in the future. There have been huge gains in knowledge of the strengths and weaknesses of national administrative records systems to support various Census Bureau activities, in the capacity for large scale data processing, data standardization, record linkage, file unduplicaton, and Social Security Number search and verification that will have benefits throughout the Census Bureau.

Research agenda for other Census Bureau programs

A research agenda for other Census Bureau programs could include:

- Testing the use of Statistical Administrative Records System as a contributor to total population and age/race/sex/Hispanic origin intercensal estimates.

- Testing the use of Statistical Administrative Records System data for improving noninterview weights in ongoing surveys.

- Testing the use of Statistical Administrative Records System as a tool to support small area income and poverty estimates.

- Continuing to test the use of Administrative Records databases for Social Security Number validation and search strategies.

- Continuing to improve our record linkage capabilities (for example, linking Current Population Survey addresses and persons to comparable Decennial Census addresses and persons), both in terms of improvements and search strategy improvements.

Intentionally Blank

1. BACKGROUND

1.1 Introduction

The Administrative Records Experiment 2000 (AREX 2000) was an experiment in two areas of the country designed to gain information regarding the feasibility of conducting an administrative records census (ARC), or the use of administrative records in support of conventional decennial census processes. The first experiment of its kind, AREX 2000 was part of the Census 2000 Testing, Experimentation, and Evaluation Program. The focus of this program was to measure the effectiveness of new techniques, methodologies, and technologies for decennial census enumeration.

Interest in taking a decennial census by administrative records dates back at least as far as a proposal by Alvey and Scheuren (1982) wherein records from the Internal Revenue Service (IRS) along with those of several other agencies might form the core of an administrative records census. Knott (1991) identified two basic ARC models: (1) the Top-down model that assembles administrative records from a number of sources, unduplicates them, assigns geographic codes, and counts the results; and (2) the Bottom-up model that matches administrative records to a master address file, fills the addresses with individuals, resolves gaps and inconsistencies address by address, and counts the results. There have been a number of other calls for ARC research — see for example Myrskyla 1991; Myrskyla, Taeuber and Knott 1996; Czajka, Moreno and Shirm 1997; Bye 1997. All of the proposals fit either the Top-down or Bottom-up model described here.

Knott also suggested a composite Top-down/Bottom-up model that would unduplicate administrative records using the Social Security Number (SSN), then match the address file, and proceed as in the Bottom-up approach. In overall concept, AREX 2000 most closely resembles this composite approach.

More recently, direct use of administrative records in support of decennial applications was cited in several proposals during the Census 2000 debates on sampling for Nonresponse Followup (NRFU). The proposals ranged from direct substitution of administrative data for non-responding households (Zanutto, 1996; Zanutto and Zaslavsky, 1996; 1997; 2001) to augmenting the Master Address File development process with U.S. Postal Service address lists (Edmonston and Schultze, 1995:103). AREX 2000 provided the opportunity to explore the possibility of NRFU support.

The Administrative Records Research (ARR) staff of the Planning, Research, and Evaluation Division (PRED) performed the majority of coordination, design, file handling, and certain field operations of the experiment. Various other divisions within the Census Bureau, including Field Division, Decennial Systems and Contracts Management Office, Population Division, and Geography Division supported the ARRS staff.

Throughout this report, rather than identifying individual workgroups or teams, we shall refer to the operational decisions made in support of AREX to be those of ARRS; that is, we shall say that “ARRS decided to…” whenever a key operational decision is described, even though, of course, ARRS were not the only decision makers.

1.2 Administrative Record Census—Definition and Requirements

In the AREX, an administrative record census was defined as a process that relies primarily, but not necessarily exclusively, on administrative records to produce the population counts and content of the decennial census short form with a strong focus on apportionment and redistricting requirements. Title 13, United States Code, directs the Census Bureau to provide state population counts to the President for the apportionment of Congressional seats within nine months of Census Day. In addition to total population counts by state, the decennial census must provide counts of the voting age population (18 and over) by race and Hispanic origin for small geographic areas, currently in the form of Census blocks, as prescribed by PL 94-171 (1975) and the Voting Rights Act (1964). These data are used to construct and evaluate state and local legislative districts.

Demographically, the AREX provided date of birth, race, Hispanic origin, and sex, although the latter is not required for apportionment or redistricting purposes. Geographically, the AREX operated at the level of basic street address and corresponding Census block code. Unit numbers for multi-unit dwellings were used in certain address matching operations and one of the evaluations; but generally, household and family composition were not captured. In addition, the design did not provide for the collection of sample long form population or housing data, needs that will presumably be met in the future by the American Community Survey program. The design did assume the existence of a Master Address File and geographic coding capability similar to that available for Census 2000.

1.3 AREX Objectives

The principal objectives of AREX 2000 were twofold. The first objective was to develop and compare two methods for conducting an administrative records census, one that used only administrative records and a second that added some conventional support to the process in order to complete the enumeration. The evaluation of the results also included a comparison to Census 2000 results in the experimental sites.

The second objective was to test the potential use of administrative records data for some part of the NRFU universe, or for the unclassified universe. Addresses that fall into the unclassified status have very limited information on them—so limited, in fact, that the address occupancy status must be imputed, and, conditional on being imputed “occupied,” the entire household, including characteristics, must be imputed. In order to effectively use administrative records databases for substitution purposes; one must determine which kinds of administrative record households are most likely to yield similar demographic distributions to their corresponding census households.

Other more general objectives of the AREX included the collection of relevant information, available only in 2000, to support ongoing research and planning for administrative records use in the 2010 Census, and the comparison of an administrative records census to other potential 2010 methodologies. These evaluations and other data will provide assistance in planning major components of future decennial censuses, particularly those that have administrative records as their primary source of data.

1.4 AREX Top-down and Bottom-up Methods

Top-down

A two-phase process accomplished the AREX 2000 enumeration. The first phase involved the assembly and computer geocoding of records from a number of national administrative record systems, and unduplication of individuals within the combined systems. This was followed by two attempts to obtain and code physical addresses (clerical geocoding and request for physical address) for those that would not geocode by computer. Finally, there was a selection of "best" demographic characteristics for each individual and "best" street address within the experimental sites. Much of the computer processing for this phase was performed as part of the Statistical Administrative Records System (StARS) 1999 processing (Judson, 2000; Farber and Leggieri, 2002). As such, StARS 1999 was an integral part of AREX 2000 design.

One can think about the results of the Top-down process in two ways. First, counting the population at this point provides, in effect, an administrative-records-only census. That is, the enumeration includes only those individuals found in the administrative records, and there is no other support for the census outside of activities related to geocoding. AREX 2000 provides population counts from the Top-down phase so that the efficacy of an administrative-records-only census can be assessed.

However, one might expect an enumeration that used only administrative records to be substantially incomplete. Therefore, a second way to think about the Top-down process is as a substitute for an initial mailout in the context of a more conventional census that would include additional support for the enumeration.

Bottom-up

The fundamental difference between the Bottom-up method and the Top-down method is the Bottom-up method matches administrative records addresses to a separately developed "frame" of addresses, and based on this match, performs additional operations. In this experiment, an extract of the Census Bureau's Master Address File (MAF) served as the frame[1].

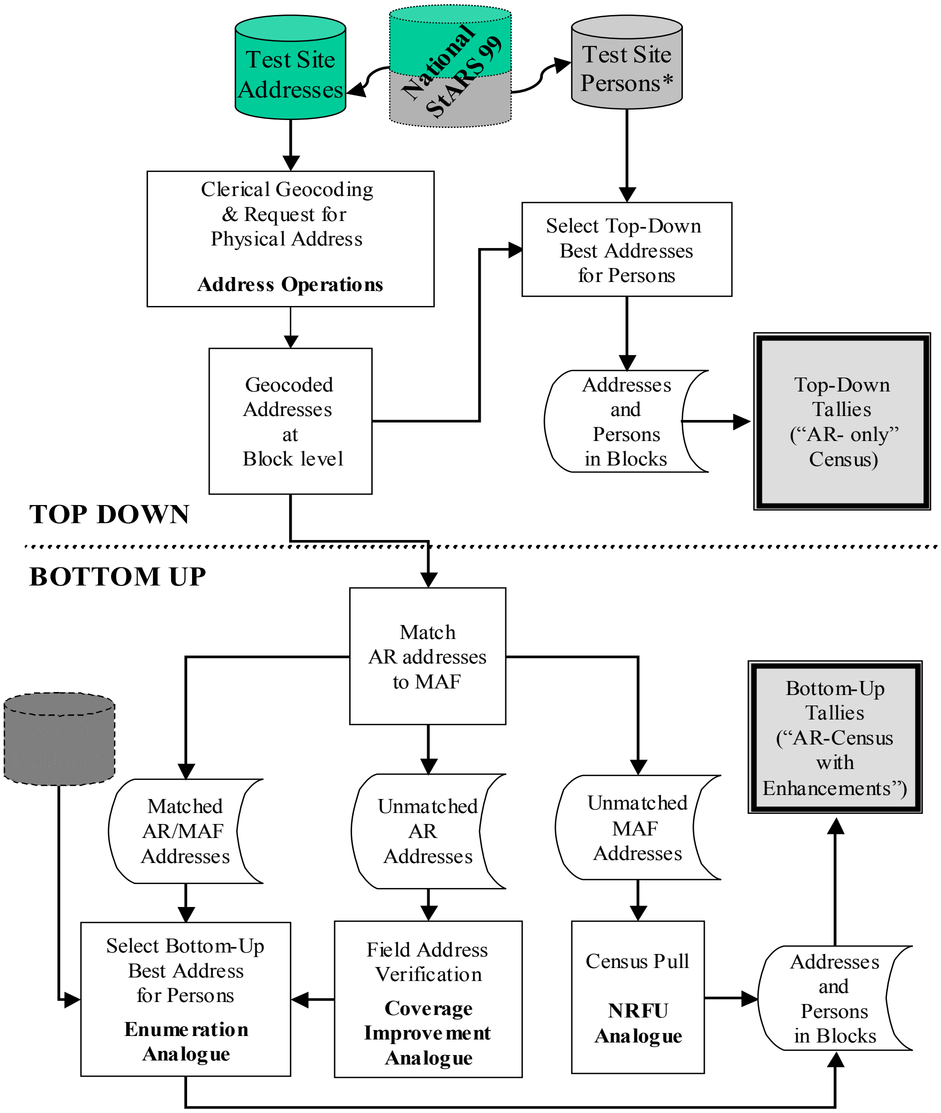
The second phase of the AREX 2000 design was an attempt to complete the administrative-records-only enumeration by the correction of errors in administrative records addresses through address verification (a coverage improvement analogue) and by adding persons missed in the administrative records (a NRFU analogue). This phase began by matching the addresses found in the Top-down process to the MAF in order to

[1] In this report, we use the term "MAF" generically. Our operations were based on extracts from the Decennial Master Address File (DMAF).

assess their validity and to identify those MAF addresses for which no administrative records were found. A field address review (FAV) was used to verify non-matched administrative records addresses, and invalid administrative records addresses were excluded from the Bottom-up selection of best address. In design, non-matched MAF addresses would be canvassed in order to enumerate persons at addresses not found in the administrative records systems. In the AREX, such a canvassing was simulated by adding those persons found in the Census 2000 at the unmatched addresses to the adjusted administrative-records-only counts, thus completing the enumeration. Accomplishing the AREX as part of the Census 2000 obviated the need to mount a separate field operation to canvass unmatched MAF addresses.

Considering the Top-down and Bottom-up processes as part of one overall design, AREX can be thought of as a prototype for a more or less conventional census with the initial mailout replaced by a Top-down administrative records enumeration. Figure 1 below, provides a conceptual overview of the experiment for enumerating the population tested during the AREX. A more detailed description of data processing flows can be found in Attachment 1. The graphical description presented here is intended to convey the concept of both AREX methods when viewed in terms of the Bottom-up method as a follow-on process to the Top-down method.

Figure 1. Summary Diagram of AREX 2000 Design


Test Site Addresses
National StARS 99
Test Site Persons*
Clerical Geocoding & Request for Physical Address
Address Operations
Select Top-Down Best Addresses for Persons
Geocoded Addresses at Block level
Addresses and Persons in Blocks
Top-Down Tallies ("AR- only" Census)
TOP DOWN
BOTTOM UP
Match AR addresses to MAF
Bottom-Up Tallies ("AR-Census with Enhancements")
Matched AR/MAF Addresses
Unmatched AR Addresses
Unmatched MAF Addresses
Select Bottom-Up Best Address for Persons
Enumeration Analogue
Field Address Verification
Coverage Improvement Analogue
Census Pull
NRFU Analogue
Addresses and Persons in Blocks

1.5 Experimental Sites

The experiment was set up to include geographic areas that include both difficult and easy to enumerate populations. Two sites were selected believed to have approximately one million housing units and a population of approximately two million persons. One site included Baltimore City and Baltimore County, Maryland. The other site included Douglas, El Paso, and Jefferson Counties, Colorado. The sites provided a mix of characteristics needed to assess the difficulties that might arise in conducting an administrative records census. Approximately one half of the test housing units was selected based on criteria assumed to be easy-to-capture in an administrative records census (for example, areas having a preponderance of city style addresses, single family housing units, older and less mobile populations), and the other half was selected based on criteria assumed to be hard to capture (the converse).

1.6 AREX Source Files

The administrative records for AREX were drawn from the StARS 1999 database. There were six national-level source files selected for inclusion in StARS. A later section of this document describes the source files in detail. The files were chosen to provide the broadest coverage possible of the U.S. population, and to compensate for the weaknesses or lack of coverage of a given segment of the population inherent in any one-source file. See Section 2 for a description of the source file characteristics.

Timing

An important limitation for the AREX is the gap between the reference period for data contained in each source file and the point-in-time reference of April 1, 2000 for the Census. The time lag has an impact on both population coverage—births, deaths, immigration and emigration—and geographic location—housing extant, and geographic mobility. As an example, both IRS files include data for tax year 1998 with an expected current address as of tax filing time close to April 15, 1999. Note, however, that the IRS 1040 file only provided persons in the tax unit as of December 31, 1998. The pertinent reference dates for each of the files are provided in Section 2.

State, Local and Commercial Files

ARRS decided not to use state and local files[2] and commercially available databases[3] in the AREX 2000 experiment. Statistical evidence is limited, but various reports from ARRS indicated that state and local files come in an extremely diverse variety of forms, with equally diverse record layouts and content (for historical information, see Sweet, 1997; Buser, Huang, Kim, and Marquis, 1998; and other papers in the Administrative Records Memorandum Series). Furthermore, ARRS reported that it was quite time-consuming and intricate to develop the interagency contractual arrangements necessary to use state and local files. Public opinion results such as Singer and Miller (1992), Aguirre International (1995), and Gellman (1997), convinced ARRS that public sensitivity to the idea of linking commercial databases with government databases (other than for address processing) would be too great, and that such a linkage would be unwise.

Census Numident

An additional, and critical, file used in creation of the StARS database was the Census Numident file. For the AREX, it was the source of most of the demographic characteristics and some of the death data. Detailed discussion regarding the creation and use of the Census Numident may also be found in Section 2.

[2] Such as state and local tax returns, drivers license files, local utilities, assessor's records, and the like.

[3] Such as commercially available mailing lists, credit card databases, and the like.

1.7 AREX Evaluations

This report is a consolidation of four evaluations of AREX 2000 that have been prepared by ARRS staff.

The **Process Evaluation** (Berning and Cook, 2002) documents and analyzes selected components or processes of the Top-down and Bottom-up methods in order to identify errors or deficiencies. It is designed to catalogue the various processes by which raw administrative data became final AREX counts and attempts to identify the relative contributions of these various processes.

The **Request for Physical Address (RFPA) Evaluation** (Berning, 2002) assesses the impact of noncity-style addresses. These addresses present a significant hurdle to the use of an administrative records census on either a supplemental or substitution basis. A particular problem is the determination of residential addresses and their associated geographic block level allocation for individuals whose administrative record address is a P.O. Box or Rural Route.

The **Outcomes Evaluation** (Heimovitz, 2002) is a comparison of Top-down and Bottom-up AREX counts by county, tract, and block level counts of the total population by race, Hispanic origin, age groups and gender, with comparable decennial census counts. This evaluation is outcome rather than process oriented.

The **Household Evaluation** (Judson and Bauder, 2002) focuses on household-level comparisons between administrative records and Census 2000. It assesses the potential for NRFU substitution and unclassified imputations, and predictive capability.

1.8 Limitations of the Experiment

In order to achieve a full understanding of the AREX processes and outcomes, it is important to appreciate the context within which the experiment was carried out. The AREX was the first attempt by the Census Bureau to experiment with the use of administrative records as the foundation of a short form decennial census. Planning for the experiment did not begin until the end of 1997, which was quite late in the Census 2000 development cycle for an experiment of such complexity. The resources for the experiment were limited to a part of the Administrative Records Research Staff (ARRS) in the Planning, Research, and Evaluation Division (PRED) with the help of other decennial census staff.

Administrative records source files

A consequence of the short planning time and limited resources was a number of design and operational decisions that made the AREX 2000 enumeration process quite different from the way such an enumeration might be carried out if administrative records were to be used in some future decennial census. Chief among these differences was the decision to use StARS 1999, the national administrative records database developed by ARRS, as the source of the administrative records for the experiment. The administrative records source files for StARS 1999 neither exhausted the national-level administrative records that might have been available for AREX 2000 nor were they the timeliest. To cover the population, StARS 1999 relied primarily on tax records for 1998 received by the Internal Revenue Service (IRS) in calendar year 1999. While IRS tax records would have to be the core of any national administrative records database, the coverage deficiencies are well known--adults without tax documents, children of taxpayers with more than four dependents, and children of adults who did not have to file 1040 income tax returns. With additional time, more could have been done to obtain administrative records from Social Security Administration (SSA) and the Centers for Medicare and Medicaid Services (CMS) that might have filled these coverage gaps. However, the acquisition of data from Federal agencies is a difficult, time consuming, and sometimes expensive process involving negotiations, interagency agreements, data extract specifications, and testing and validation of the delivered products before such data can be included in a national data base.

Obtaining timelier data for the AREX 2000 would have required, in some cases, the receipt of data on a flow basis from the source agencies. Receipt of tax forms filed in calendar year 2000, would have required obtaining 1040 data from IRS on a flow basis and possibly 1099 and W-2 data from SSA as well. (See Section 2 for a discussion of the sources and timing of tax data.) Also, more timely extracts might have been obtained from other contributing agencies had there been sufficient time to make the arrangements. As will be evident in this report, the fact that the reference period for the administrative data was one or more years behind census day (April 1, 2000) was the single most important limitation to the AREX goal of testing the completeness and accuracy of an administrative records census.

Using timely administrative records data in a decennial census implies large administrative records data processing operations would be done quickly as part of the decennial enumeration. One thing learned from AREX 2000 is that such processing is technically feasible and could be accomplished with the planning time and resources that would be available for actual census operations as opposed to those typically available for small experiments.

Two experimental sites

A second major limitation imposed by lack of planning time and resources was the restriction of the experiment to five counties in two states. Although it would not have been reasonable or realistic to attempt to mount this first AREX in a representative sample of geographic areas large enough to make national estimates, additional sites would have provided more confidence that the results could be generalized beyond the sites selected. While there is much to be learned from the AREX, it is important to keep in mind that for the AREX results, descriptive statistics are generally only representative of the test sites themselves; and the modeling results, though suggestive of the relationships between administrative records outcomes and their covariates, are not definitive.

Lack of experimental variation of key design parameters

There were several AREX operations relating to address processing that could have been more thoroughly evaluated with some additional structure in the experimental design. These operations involved clerical and field attempts to validate addresses and obtain block-level geocodes, clerical support for addressing matching of administrative records to the Master Address File, and the "best address" selection algorithm for the administrative records. In all cases, the objective of the evaluation would have been to assess the impact of the particular operation or algorithm on final address selection and ultimately whether the operation contributed significantly to the accuracy of the AREX enumeration.

Evaluation of the clerical and field operations, individually or in combination, would be important because they represent potential costly components were they to be implemented as part of a national administrative records census. Evaluation of the address selection algorithm would have revealed the impact of the preference of geocoded addresses over others in the algorithm. Unfortunately, the experimental design did not include factorial or fractional factorial structure permitting direct estimates of the impact of operational components, individually or in combination.

Race and Hispanic origin models

Population tallies by race and Hispanic origin are a crucial product of the short form census because of their use in drawing and evaluating political districts at and below the state level. Measurement of race and Hispanic origin is a major weakness of administrative records at the national level and any attempt to use administrative records to enumerate all or part of the population would have to find some way of improving the information available in administrative records.

In his design proposal for an administrative record census in 2010, Bye (1997) suggested building a list of SSNs annotated by race and Hispanic origin by a series of operations that would begin by matching Census 2000 to SSA's Numident and continue during the years leading up to the 2010 Census. (See also Bye and Thompson (1999). Sections 4 and 5 of this report describe activities currently underway at the Census Bureau.) Had more planning time and resources been available to the AREX, it might have been possible to incorporate race roster building into the experiment by including one or more of the 1995 and 1996 census test sites or Census 2000 Dress Rehearsal sites in the AREX (Bye, 1997).

However, such race roster building was not available to the AREX, and ARRS decided to use Numident-based national-level models to augment the race and Hispanic origin data (Bye 1998). Although using the models generally worked in aggregate counts, the use of national-level models to impute characteristics of small geographic areas has certain well-known weaknesses in that the actual findings in the smaller areas can vary substantially around the national predictions. Bye (1998) provided tabulations for states and some substate areas showing the kind of variation that could be expected when using the national models for the AREX. Bye and Thompson (1999) provided a partial solution to this problem, but an annotated Numident file is clearly a superior solution.

2. THE AREX PROCESS EVALUATION

2.1 Introduction

This section describes and evaluates the AREX enumeration processes. The process description is taken largely from the process evaluation report of Berning and Cook (2002). In this report, process descriptions have been provided separately for the Top-down and Bottom-up enumerations. The actual data processing flows were often intermingled and are provided by Berning and Cook in great detail. Concerning process evaluation, Berning and Cook focused mainly on data processing and clerical operations.

Administrative records

AREX source files

The administrative records for AREX were drawn from the StARS 1999 database. The six national-level source files selected for StARS were chosen to provide the broadest coverage possible of the U.S. population. At a minimum, the files had to have for each record, a name, Social Security Number (SSN), and street address.

The national level files that contributed to the StARS 1999 database and therefore to AREX 2000, were:

Internal Revenue Service (IRS) Tax Year 1998 Individual Master File (IMF 1040),

IRS Tax Year 1998 Information Returns Master File (IRMF W-2 / 1099),

Department of Housing and Urban Development (HUD) 1999 Tenant Rental Assistance Certification System (TRACS) File,

Center for Medicare and Medicaid Services (CMS) 1999 Medicare Enrollment Database (MEDB) File,

Indian Health Services (IHS) 1999 Patient Registration System File, and

Selective Service System (SSS) 1999 Registration File.

The following table displays the primary reason each file was included in the StARS database and the approximate number of input records associated with each.

Table 1. Source File Characteristics

File	Targeted Population Segment	~ # of Address Records	~ # of Person Records
IRS 1040	Taxpayer and other members of the reporting unit with current address	120 million	243 million
IRS W2/1099	Persons with taxable income who might not have filed tax returns	598 million	556 million
HUD TRACS	Low income housing population (possible non-taxpayers)	3.3 million	3.3 million
Medicare File	Elderly population (possible non-taxpayers)	57 million	57 million
IHS File	Native American population (possible non-taxpayers)	3.1 million	3.1 million
SSS File	Young male population (possible non-taxpayers)	14.4 million	13.1 million
	Total	795 million	875 million

Notes: The variance between the number of address records and person records within the input source files is a result of the following source file characteristics:

1. Each IRS 1040 input record may reflect up to six persons (primary filer, secondary, and dependents).
2. Each SSS input record may reflect two addresses - defined as current and/or permanent address.
3. The IRS W-2/1099 file undergoes a preliminary unduplication and clean-up process prior to the initial file edit process.

Timing

An important limitation for the AREX was the gap between the reference period for data contained in each source file and the point-in-time reference of April 1, 2000 for the Census. The gap had an impact on both population coverage (births, deaths, immigration and emigration) and geographic location (housing extant, and geographic mobility). As an example, the IRS 1040 file included data for tax year 1998 with an expected current address as of tax filing time close to April 15, 1999, but provided only persons in the tax unit as of December 31, 1998.

The following table displays the reference periods of the files available. Generally, the reference periods are about one year prior to the day of Census 2000.

Table 2. Reference Dates of Source Files

Source File	Cut-off Date	Requested Cut Date	Universe
Indian Health Service	04/01/99	04/01/99	All persons alive at cut-off date
Selective Service	Note 2	04/01/99	Males between the age of 18 – 25
HUD TRACS	04/01/99	04/01/99	All persons on file as of cut-off date
Medicare	Note 3	04/01/99	All persons alive at cut-off date
IRS 1040	12/98	Note 1	Individual tax returns for tax year 1998
IRS W-2 / 1099	12/98	04/01/99	Forms W-2 and all 1099 forms tax year 1998

Notes:

1. File Cut date is for posting cycle weeks 1-39 only for IRS 1040, and weeks 1-41 for IRS 1099 files. Weeks 40-52 (and 42-52 respectively) were not included in StARS '99. This file reflects the most current address on file for the taxpayer. It could be an address that has been updated since the 1998 tax return was posted.

2. Cut-off date is same as dates used to define universe: persons born after April 2, 1972 and on (or before) April 1, 1980.

3. Universe also defined as persons with a death date of 12/31/1989 or later.

Census Numident

An additional, and critical, file used in creation of the StARS database was the Census Numident file. For the AREX, it was the source of most of the demographic characteristics and some of the death data.

The Census Numident was created by ARRS for the primary purpose of validating Social Security Numbers (SSNs) used in the processing of administrative records and supplying demographic variables missing from source files. The Census Numident is an edited version of the Social Security Administration's (SSA) Numerical Identification (Numident) File. The SSA Numident file is the numerically ordered master file of assigned Social Security Numbers (SSN) that may contain up to 300 entries for each SSN record, although on average contains two records per SSN. Each entry represents an initial application for a SSN or an addition or change (referred to as a transaction) to the information pertaining to a given SSN. The SSA Numident contains all transactions (and therefore, multiple entries) ever recorded against a single SSN. The SSA Numident available for StARS 1999 reflected all transactions through December 1998.

The Census Numident was designed to collapse the SSA Numident entries to reflect "one best record" for each SSN containing the "best" demographic data for each SSN on the file. Following edit, unduplication, and selection of best demographics, the SSA Numident file of nearly 677 million records was reduced to just over 396 million records that comprise the Census Numident file.

2.2 Top-down enumeration

Dual stream process

The goal of the Top-down process was to use administrative records to identify individuals residing at geocoded addresses in the AREX test sites and to construct a data record for each individual that contained demographic data (age, gender, race and Hispanic origin) corresponding as closely as possible to census short form data. To achieve this goal, a "dual-stream" processing approach was adopted. One processing stream concerned the development of a unique record for each individual with best demographics. The second stream involved development of an unduplicated set of addresses, geocoded to the block level. In the end, persons and addresses were brought together, and a best address was selected for each person to complete the Top-down enumeration.

The following sections provide a brief description of the AREX Top-down data processing steps. Much of the work was accomplished in the development of StARS 1999 itself, but there were some differences in demographic and address selection rules. More detail is given in Berning and Cook (2002).

Top-down Person processing consisted of three main steps.

1. File edits for person data,

2. SSN Verification of person records,

3. Unduplication of person records, and creation of the Person Characteristics File (PCF) that contained the "best" demographic characteristics for each person record.

Models were used to generate “best” demographic characteristics. Details about the models can be found in Bye (1998, race/Hispanic origin)[4], and Thompson (1999, gender)[5]. In general, a person’s modeled race or gender was used only in the case where no race appeared on any administrative record, including the Numident**.** In the case of gender, the model was rarely used since the Numident reported gender more than 99 percent of the time. For Race, the model was used when the Numident race was shown as “Other” or “Unknown” or “Hispanic,” and no other administrative record provided it. The vast majority of cases with unknown race were either children whose applications for SSNs were processed via SSA’s enumeration-at-birth program, which was started in the mid-1980s, or older persons who had applied for Social Security benefits prior to SSA’s development of the electronic Numident in the mid-1970s.

For Hispanic Origin, the model was used for all cases for which neither the Numident nor any of the other administrative records indicated Hispanic origin. Because the Numident did not capture Hispanic origin prior to 1980, the model was used for well over 90 percent of the cases. The following table shows the extent of Race and Hispanic Origin imputation for the individuals included in the Top-down AREX enumeration.

Table 3. Percent of cases with imputed Race or Hispanic Origin by age and County

	Imputed Race		**Imputed Hispanic Origin**	
County	<18	18 and over	<18	18 and over
Baltimore City	40.7	2.8	99.4	96.7
Baltimore County	49.3	4.0	99.2	98.5
Douglas, CO	58.2	6.0	98.3	97.7
El Paso, CO	52.2	9.0	93.9	93.9
Jefferson, CO	54.8	8.0	94.7	95.2

[4] The Race and Hispanic Origin models were developed using Numident data and Spanish and Asian name lists. The principal variables in the prediction equations were: (1) race or Hispanic origin as it appeared in the Numident, (2) place of birth, (3) Spanish and Asian surname indicators for the SSN holder and parents’ surnames, and (4) indicator field in the Indian Health Service file. The Race and Hispanic Origin models were originally developed to augment race and Hispanic origin information in the Numident.

[5] The gender model was based on the strength of association between first and middle names and reported sex. Look-up tables created for common names, uncommon names, name-gender proportions, and gender model parameters were created and a final gender probability assigned after the four look-up tables were created and run against each input record.

Top-down Address Processing

Top-down Address Processing consisted of four main steps.

1. File edits for address data.
2. Code-1 processing and computer geocoding the address records.
3. Manual geocoding for addresses not coded by computer.
4. Creation of Master Housing File for administrative record addresses.

The creation of the Master Housing File for administrative record addresses was the final step in the address processing before the addresses were relinked with the person records. This step had two main objectives. First there was an attempt to identify commercial addresses in the files. Second, there was a final attempt to unduplicate the addresses prior to the application of address selection rules.

Table 4. StARS 1999 and AREX Test Site Geocoding Tallies

	# Input Records to Geocoding	**# of Records Geocoded**	**Percent Geocoded**
StARS National Address File	147,346,145	108,032,169	73.3
Maryland subset of StARS National File	725,108	626,247	86.4
Colorado subset of StARS National File	624,248	498,783	79.9

Clerical Geocoding and Request for Physical Address (RFPA)

Addresses that cannot be geocoded by computer generally fall into three categories: (1) city style addresses; (2) P.O. Box and non-city style addresses (rural route/box number); or (3) addresses that are so fragmented that they cannot be classified. Procedures for attempting to obtain geocodes for the first two classes of addresses are described below. Seriously fragmented addresses are discarded at this point.

Master Address File Geocoding Office Resolution (MAFGOR)

MAFGOR was an existing operational capability within the Regional Census Centers (RCC) to provide clerical geocoding for the Decennial Master Address File as part of

Census 2000. Addresses identified by ZIP code as being potentially in the AREX sites but not geocoded by computer were sent to the Philadelphia RCC (79,307) and the Denver RCC (83,841). These two RCCs attempted to clerically geocode these addresses using trained staff, reference materials, and maps. The clerical geocoding added about 3 percent to the total number of addresses coded.

Request for Physical Address (RFPA)

P.O. Box and rural route/box number addresses pose a special challenge for geocoding. The P.O. Box address does not refer to a physical location and the non-city style addresses often do not precisely identify the housing unit location. The RFPA was an attempt to collect physical addresses (house number and street name) for persons receiving mail at these potential test site addresses. Major components of the operation were to:

- Create an address file from administrative records where the mailing address was a Post Office Box or noncity-style address.
- Design and mail a form requesting physical address information.
- Have the RCCs attempt to clerically geocode the physical addresses of the returned forms to state, county and block.
- Key addresses and geocode information to a file for further analysis.

The mailing was sent to 58,151 addresses associated with 138,653 individuals. For a number of reasons, the response rate to the mailout was only about 20 percent of which about 86 percent (9,431 physical addresses) were geocoded, 8,090 to an AREX test site county. The coded addresses were to have been added to the address lists prior to AREX address selection. However, because of the small number of persons that would have been potentially added to the enumeration or for whom addresses might have changed, these addresses were not incorporated into the AREX address file. As indicated above, the RFPA was the subject of a special evaluation. More can be found in Berning (2002).

Table 5 provides a summary of Top-down address coding. Note that only about 3,000 addresses were too fragmented to be eligible for either MAFGOR or RFPA.

Table 5. AREX Administrative Record Address Geocoding Results

State	Addresses in Test Sites	TIGER Coded	Not TIGER Coded	Eligible for MAFGOR	Coded by MAFGOR
Maryland	725,108	626,247	98,861	79,307	21,542
Colorado	624,248	498,783	125,465	83,841	28,030
Total	1,349,356	1,125,030	224,326	163,148	49,572

	Not Eligible for MAFGOR	Eligible for RFPA	Returned with Useable Info.	Coded to AREX Site	Not Eligible for MAFGOR/ RFPA
Maryland	19,544	18,694	3,538	1,939	860
Colorado	41,624	39,457	8,145	6,151	2,167
Total	61,168	58,151	11,683	8,090	3,027

AREX Master Housing File

The AREX Master Housing File (MHF) contained an unduplicated set of non-commercial address records that was linked with the person records prior to the application of the best address selection algorithm.

AREX Top-down composite person records (CPR)

At this point in the AREX "dual stream" process, address and person data were brought together in preparation for creation of the Composite Person record. There were two principal tasks. First, individuals potentially in the AREX test sites were identified. Then, the best address was selected for these persons. If the best address was in the test site, then the individual became part of the Top-down enumeration.

The development of the AREX person universe began with the national databases of persons and addresses described in the previous sections. First, all persons ever associated with an AREX address were included in a file of potential AREX persons. Next, all of the addresses associated with these persons--addresses both in and outside of the test site--were assembled and subjected to the following selection algorithm.

° Select geocoded addresses over non-geocoded addresses.

° Select the highest HUID category available.

° Select a non-proxy address over an address with a proxy.

° Select a non-commercial address over a commercial address.

- Select the address based on source file priority as follows:
  - IRS 1040 record
  - Medicare record
  - Indian Health Service record
  - IRS 1099 record
  - Selective Service record
  - HUD TRACs record
  - Select the most recent record based on the administrative record cycle dates.
  - Select the first record read-in to the processing array for output to the CPR.

  If the best address for any person record from among the AREX person universe file was determined not to be within the AREX test site, the person record was flagged "out of scope" to ensure the person was not counted in the population tallies for the AREX test site.

Top-down process results

The composite person record represents the completion of the Top-down process for the AREX 2000 experiment. Prior to tabulation, a final match of the AREX addresses was made to the Decennial Master Address File (DMAF) for the purpose of transforming the collection geography to tabulation geography[6]. Because the AREX addresses were initially geocoded to collection geography, it was necessary to translate the collection geographic codes into the tabulation geographic codes so that the comparisons to Census 2000 tabulations could be made.

The tallies for the top down method are shown in the following table.

[6] The taking of the census spans approximately a two year period, including the address list building phase. The geographic framework going into the census is called collection geography. Prior to tabulation of the final Census counts, changes must be incorporated to reflect boundaries in effect on January 1, 1999. This final geographic framework is called "tabulation" geography.

Table 6. Top-down Population Tallies

Test Site County	AREX Population	Census 2000 Population	Percent of Census Population
Baltimore City Maryland	570,648	651,154	88%
Under 18	134,471	161,353	83%
18 and over	436,127	489,801	89%
Baltimore County Maryland	696,183	754,292	92%
Under 18	146,012	178,363	82%
18 and over	550,086	575,929	96%
Douglas County Colorado	148,270	175,766	84%
Under 18	40,085	55,477	72%
18 and over	108,165	120,289	90%
El Paso County Colorado	456,891	516,929	88%
Under 18	110,504	142,480	78%
18 and over	346,322	374,449	92%
Jefferson County Colorado	473,495	527,056	90%
Under 18	101,535	133,486	76%
18 and over	371,894	393,570	94%

The counts by age showed the expected results. Generally, administrative records undercounted the population; but coverage of adults (89 percent - 96 percent) was much better than children (72 percent - 83 percent). There is an evaluation of the administrative records data sources and Top-down processing tasks in Bye 2002.

2.3 Bottom-up enumeration

The weaknesses of the Top-down process as exhibited above were not unexpected. In fact, most historical proposals for an administrative records census recognized that additional operations, beyond tallies of administrative records, would have to be performed for a complete enumeration to be obtained.

The Bottom-up phase of the AREX 2000 design was an attempt to complete the administrative-records-only enumeration by adding persons missed in the administrative records, a process analogous to a conventional nonresponse followup (NRFU). There was also an attempt to correct Top-down enumeration errors by removal of invalid administrative records addresses prior to best address selection. A valid address was defined as one that matched the DMAF or was deemed valid after a field address review. There was no provision for correcting enumerations at households with valid administrative records addresses. Non-matched DMAF addresses were canvassed in order to enumerate persons at addresses not found among the validated administrative records addresses. In the AREX, the canvassing was simulated by adding those persons found in Census 2000 at the unmatched addresses to the adjusted administrative-records-only counts, thus completing the enumeration. This phase of the AREX was designated as Bottom-up because it started with a known list of residential addresses (in this case the DMAF), matched the administrative records addresses to such a list, and reconciled any non-matched cases.

The Bottom-up operational components of AREX were conducted on records contained within the five test site counties. These operations consisted of:

- Computer matching AREX addresses to the DMAF.
- Clerical review of unmatched administrative records addresses.
- Field Address Verification of unmatched administrative record addresses
- Address re-selection.
- Census Pull, the simulated NRFU.
- Bottom-up enumeration.

Matching AREX records to the DMAF

<u>The DMAF Computer Match</u>

The objective of the computer match operation was to determine the extent and nature of agreement between addresses from administrative records source files and eligible addresses from the Census Bureau's Decennial Master Address File (DMAF). To most accurately match the addresses, the AREX addresses were limited to those, which were geocoded, or with a standardized street name, a standardized property description or both. Excluded from the matching process were non-standardized addresses, standardized post office or box addresses, standardized post office and rural route addresses and undefined addresses. Table 7 shows the administrative records addresses and the DMAF addresses eligible for the computer match.

Table 7. Addresses eligible for the match to the DMAF

	Addresses from Administrative Records			Unduplicated DMAF Addresses
		TIGER/MAFGOR		
Test Site	**Total**	**Coded**	**Non-coded**	
Maryland	656,073	647,789	8,284	650,109
Colorado	531,382	526,813	4,569	526,018
Total	1,187,455	1,174,602	12,853	1,176,127

The matching process used AutoMatch, a commercial software package that applies probabilistic record linkage techniques. The final results were divided into matches; possible matches; non-matches and matches to duplicate DMAF addresses. Table 8 shows the results of the computer match for the administrative records addresses.

Table 8. Computer Match Results -- Administrative Records Counts

Test Site	# Records to Computer Match	# of Addresses Matched	% of Addresses Matched	Possible Matched Records	Non-Matched Records	Duplicate Matches
Maryland	656,073	525,234	80%	2,134	128,286	419
Colorado	531,382	432,140	81%	9,430	88,586	555
Total	1,187,455	957,374	81%	11,564	216,872	974

Some administrative records matched to more than one address in the DMAF, each of which might have had subtle differences. When this occurred, addresses were flagged as having duplicate matches. The duplicates were resolved later in the AREX operation where the best address was determined based on pre-defined criteria.

Clerical Review of Unmatched Administrative Records Addresses

Following the computer match, the staff at the National Processing Center conducted a clerical review.

The results of the clerical review are shown in Table 9.

Table 9. Clerical Review Match Results

Test Site	Records sent to Computer Match	Matched Records before Clerical Review	Matched Records after Clerical Review	% of Matched Records matched by Clerical Review
Maryland	656,073	525,234	543,811	3%
Colorado	531,382	432,140	459,753	5%
Total	1,187,455	957,374	1,003,564	4%

Field address verification (FAV)

The Field Address Verification operation was implemented to check the validity of addresses that remained unmatched to the DMAF following the computer matching and clerical review. The purposes of the FAV were to:

- Verify the physical existence or nonexistence of non-matched AREX 2000 Test Site addresses.
- Correct erroneous address field values.
- Identify addresses meeting unique conditions such as being a duplicate of another address.

The original plan called for a review of 100 percent of the unmatched addresses by census field staff, but the plan was changed to have only a sample of addresses reviewed by Census Bureau Headquarters volunteers. The results from the sample were used to estimate a regression equation giving the probability of a valid address. The equation was then used to impute validity or lack thereof to the non-sample addresses.

Sample design

After the computer phase of address matching, the universe of addresses eligible for Field Address Verification was first restricted to geocoded, city-style addresses within the AREX 2000 test site counties. The universe was further restricted to exclude some AREX 2000 test site ZIP codes that belonged to three colleges, a medical center, and an Air Force base in the belief that few or no residential addresses existed in those areas.

With the redesign of the FAV operation, the addresses to be verified were based on a stratified cluster (Census block) sample of unmatched, city style addresses. The sample consisted of 112 blocks per AREX county and resulted in 6,644 addresses being flagged as part of the FAV sample (table 10).

Table 10. Selection of FAV Addresses

Test Site State	Number of FAV Eligible Addresses	Number of Addresses Selected for FAV Sample
Maryland	96,202	2,914
Colorado	57,333	3,730
Total	153,535	6,644

After the fieldwork was completed and the results keyed, PRED staff then reviewed each of the listing pages and annotated a 5-digit status code on the page. The code categorized the type of activity about the address that was shown on the listing page and the validity of the address. In some instances, addresses were determined to be valid as listed (without changes). In other cases, corrections were made to the address to make the address valid. Table 11 provides the FAV sample results.

Table 11. FAV Sample Results

Test Site	Number of Addresses Sampled	Percent Valid	Percent Valid as Listed	Percent Valid After Lister Corrections
Maryland	2,914	38%	13%	25%
Colorado	3,730	41%	7%	34%
Total	6,644	40%	10%	30%

Of particular interest in this table are the percentages of addresses determined to be valid as listed. Because these addresses did not match the DMAF even after clerical review, it is possible that the DMAF was incomplete. However, this may also reflect residual difficulties in the matching process.

Imputing validity to non-sample addresses

The FAV sample cases were used to estimate a logistic regression model, logit(P(y=1|x))= xβ. In this equation, the outcome measure y = 1 if the address was valid, y = 0 otherwise. The predictor variables, x, represented (1) characteristics of the administrative record addresses as possibly modified by the FAV review, (2) DMAF block size of the DMAF address to which the administrative record partially match, and

(3) the nature of the partial match. Generally, administrative records addresses were found more likely to be valid if they were not commercial, were found in multiple administrative record source files, had no unit identifier, and matched a DMAF address or addresses (by state, county, zip, street name, and street name suffix) for which there were no unit identifiers (i.e., it or they appeared to be a single-family dwelling or dwellings), and were located in blocks in which the DMAF indicated a fairly large number of addresses. The overall probability of misclassification--the probability that an address was not valid times the probability of a false positive plus the probability that an address was valid times the probability of false negative-- was estimated to be 0.32. (A detailed discussion of the model and regression results can be found in Bye, 2002.)

Validity or lack thereof was imputed for all FAV eligible addresses that were not part of the sample by using the regression equation to calculate the probability that the address was valid, and comparing this value to a random number drawn from a uniform distribution between 0 and 1. If the random number was less than or equal to the predicted probability, the address was deemed to be valid for AREX Bottom-up address selection purposes.

The net results of the Bottom-up administrative records addresses processing -- match of AREX addresses to the DMAF and the subsequent FAV -- are given in the following table.

Table 12. Bottom-up Administrative Records Address Processing

Test Site	Addresses sent to DMAF Computer Match	Matched Addresses after DMAF Computer Match and Clerical Review	Non-matched Addresses	FAV Eligible Addresses	Number of valid Addresses of those eligible for FAV (FAV sample or imputed)
Maryland	656,273	543,881	112,392	96,202	36,661
Colorado	531,382	459,753	71,629	57,333	23,310
Total	1,187,655	1,003,634	184,021	153,535	59,971

As a result of the FAV operations, 93,382 (153,535 - 59,971) of the FAV-eligible administrative records addresses were found to be invalid, and were not eligible for the Bottom-up address selection. Note, however, that unmatched addresses not eligible for the FAV remained in the Bottom-up address pool as a possible Bottom-up address.

Second DMAF match

A second match of the AREX addresses was made to the DMAF for the purpose of transforming the collection geography to tabulation geography. Because the AREX addresses were initially geocoded to collection geography, it was necessary to translate the collection geographic codes into the tabulation geographic codes so that the comparisons to Census 2000 tabulations could be made. In general, the difference between collection blocks and tabulation blocks was that some collection blocks were split in final decennial census tallies.

The contents of the DMAF were not stationary between the first and second match. There were a number of problems with duplicate MAFIDs for different addresses and multiple MAFIDs for the same address in both DMAF matches. Sometimes administrative records that matched the first time did not match the second. In these cases, if the original collection block was split into more than one tabulation block, then the address was statistically allocated to a tabulation block. (The second DMAF match also had an impact on Top-down block assignments.)

Bottom-up address selection and composite person records

For an address to be considered eligible for Bottom-up selection, the following conditions had to be met after the rematch to the DMAF:

1. The address had to have a Census tabulation block code.
2. The address could not have been identified as a commercial address during the FAV.
3. The address had to be either non-FAV eligible, a FAV sample address that was found to be valid during field review, deemed valid based on FAV imputation, or valid based on matching during the rematch to the DMAF.

Once the pool of eligible addresses was identified and linked to the unduplicated list of AREX persons, the address selection operations were similar to the top down selection and identification of persons in administrative records who were eligible for the Bottom-up enumeration were similar to those procedures used for Top-down selection. Generally, all the addresses associated with an individual were assembled and subjected to the address selection rules to obtain the “best” address for each individual in the administrative record source files.

A possible outcome of the address selection process was that no persons remained at valid addresses in the AREX test sites. That is, although one or more persons were originally associated with the administrative record address, best address selection resulted in all persons at the address being assigned to another test site address or to an address outside the test site. These addresses were designated as AREX vacant addresses; there were 179,523 such addresses.

Simulated NRFU -- the Census Pull

A principle feature of the Bottom-up process was to complete the enumeration by adding persons at test site addresses not found in administrative records. Presumably this would have been accomplished by some sort of mailout/mailback procedure or face-to-face interviews or both. This can be considered as an analogue to a conventional nonresponse followup; albeit in this case, the "nonresponse" is to the initial administrative records enumeration.

For the AREX, the NRFU analogue was simulated by including persons found in the Census 2000 Hundred Percent Detail File (HDF) at addresses that were not found among the administrative records addresses, occupied or vacant. These were persons enumerated in Census 2000, and the assumption was that they would have been counted in the AREX had some sort of followup been instituted. The process of including persons from the Census 2000 HDF was referred to as the Census Pull.

Table 13 shows the number of Census Pull addresses and persons included in the Bottom-up enumeration.

Table 13. Census Pull Results

State	Census 2000 Addresses	Census Pull Addresses	Census Pull Persons
Maryland	615,323	97,460	185,868
Colorado	478,701	55,319	126,558
Total	1,094,204	152,779	312,426

Of the Census Pull addresses, 35,591 were vacant Census 2000 addresses.

Revised race imputation for children under 18

Instead of using the race model to impute race for children under age 18 with unknown race in administrative records as was done in the Top-down process, an alternative imputation method was used. The source of most children in the AREX was the IRS 1040 file, which generally provided primary and secondary tax payers and up to four dependents in each tax unit. Children under 18 with unknown race, who could be associated with a tax unit, were assigned the race of the primary taxpayer.

Bottom-up results

<u>Overall enumeration</u>

The AREX Bottom-up enumeration results are shown in **Table 14**. As expected, the coverage is much improved compared to the Top-down counts, and is largely due to the completion of the Top-down enumeration by the Census Pull. Specifically, the Bottom-up coverage of children (81 percent - 94 percent across the test sites) is substantially better than the Top-down (72 percent - 83 percent). Adults in the Bottom-up are more or less uniformly overcounted (102 percent - 104 percent). The overcount of adults most likely is due to unaccounted for deaths in the previous 12 months, handling of special populations, and failure to unduplicate persons after the Census Pull (discussed later in the report). Of course, the latter means that there is some duplication for the children as well.

Table 14. Bottom Up Method Population Tallies

Test Site County	AREX Population	Census Population	% of Census Population
Baltimore City Maryland	661,561	651,154	102%
Under 18	151,411	161,353	94%
18 and over	510,109	489,801	104%
Baltimore County Maryland	745,893	754,292	99%
Under 18	154,500	178,363	87%
18 and over	591,313	575,929	103%
Douglas County Colorado	170,102	175,766	97%
Under 18	46,394	55,477	84%
18 and over	123,689	120,289	103%
El Paso County Colorado	509,597	516,929	99%
Under 18	121,647	142,480	85%
18 and over	387,888	374,449	104%
Jefferson County Colorado	508,254	527,056	96%
Under 18	108,618	133,486	81%
18 and over	399,575	393,570	102%

Net effect of Bottom-up processes on administrative records tallies

Two of the Bottom-up operations entailed an attempt to improve the administrative records addresses prior to Bottom-up "best" address selection: (1) the initial match to the DMAF and its followup clerical review, and (2) the FAV[7]. The impact of these operations on the administrative records part of the Bottom-up enumeration was threefold. First, administrative records addresses were removed from consideration if

[7] The second match to the DMAF had an impact on address selection for both the Top-down and Bottom-up and should not be considered solely a Bottom-up operation.

they did not match the DMAF, were FAV eligible but were not found or deemed to be valid by the FAV. Second, addresses that were not geocoded in the TIGER match or MAFGOR might have been coded through one or the other of these operations. Third, some of the addresses that were geocoded prior to the initial DMAF match might have received code changes.

The impact of these operations on the addresses has been discussed in the relevant sections. Table 15 provides some information on the net impact of these operations on Bottom-up person tallies from administrative records, and provides a comparison of the Top-down and Bottom-up administrative records person tallies.

Of the 2.3 million persons tallied in the Top-down enumeration, 70,031 (about 3 percent) were excluded from the Bottom-up administrative record counts. These exclusions occurred either because the only address that the persons had was rejected by the Bottom-up processes or because the only remaining addresses were outside of the AREX test sites.

For administrative records persons enumerated in both the Top-down and the Bottom-up, Table 15 provides information on the change in geographic location due to Bottom-up processes. Here, there seems to have been very little impact; over 99 percent of these persons were at the same address in both enumerations.

Table 15. Geographic Differences for Persons in both the Top-down and Bottom-up methods

	Top-down in…				
Bottom-up total	Same Address	Different Address	Different Block	Different Tract	Different County
2,275,456	2,258,441	17,015	15,129	11,847	2,363
100.0%	99.3%	0.8%	0.7%	0.5%	0.1%

Overall, the net effect of the Bottom-up operations on the administrative record tallies was quite modest.

Bottom-up evaluation

The Bottom-up evaluation focused on both operations and the goals of a decennial short form census.

AREX Bottom -up processing operations

DMAF computer match

The computer match rate between eligible AREX addresses and the DMAF was only about 80 percent. A number of factors may have contributed to the match rate level. First, there is the vintage of the administrative record addresses. Most of the AREX addresses were of 1999 vintage, one year or older than the DMAF. Destruction of housing units and changes in official address components could account for some of the non-matches.

Second, although a number of adjustments were made to the AutoMatch parameters to try to ensure optimum match rates, the clerical review following the computer match resulted in a substantial number of additional matches suggesting that there is still room for improvement in the use of matching software. The fact that most of the unmatched addresses were geocoded by TIGER or MAFGOR suggests just how difficult address matching can be.

Third, a more consistent method of address standardization should improve the overall match rate. Throughout the course of creating the StARS database and subsequent iterations of the AREX address file, the Geography Division's address standardizer was employed. The dynamic nature of the standardizer software program and the flexibility of operator control during its application most likely contributed to inconsistencies and variances that led to erroneous non-matches (and matches as well). Although difficult to quantify, the application of a fixed version of the standardizer along with prescribed operator control methodologies should improve the overall match rate during the computer matching operations. Improving the computer match rate would, in turn, reduce the number of address records requiring clerical review.

Finally, multiple MAFIDs assigned to a single address and duplicate MAFIDs assigned to multiple addresses contributed to the difficulty in classifying an address as matched, non-matched, or possibly matched. These difficulties may be due to the Census Bureau's methodology and audit trail for identification and retention of "surviving MAFIDs" on the DMAF as the DMAF changes over time. Further research needs to be done on the best formulation of DMAF extracts for administrative record matching.

Second DMAF match

Prior to the AREX enumeration, a second match to the DMAF was required to pick up "tabulation" block codes. The block codes obtained from the original TIGER match and MAFGOR operation were "collection" block codes. The difference between the codes is that some collection blocks were split as part of a final decennial census-coding scheme. The AREX needed to use the final block codes in order to facilitate comparisons between AREX and Census 2000 results. Addresses that did not match the second time and were in collection blocks that had been split by one or more tabulation blocks were statistically allocated to one of the split blocks.

Clerical review of non-matched AREX addresses

The original AREX plan called for PRED staff to do the clerical review of the unmatched and possible-matched records after the initial match of the administrative records addresses to the DMAF. However, resource constraints due to changes in FAV plans required that the review be shifted to the National Processing Center (NPC). Accordingly, PRED trained approximately 25 reviewers to evaluate the possible matches of AREX addresses against the DMAF and make a match/non match determination for the address.

Field Address Verification

To minimize the impact of the lack of experience, the listers were not used in the traditional role of assigning action codes but rather to collect information about the address for later analysis and assignment of the action code. The listers answered 11 questions about the property from which the action code (called status code in this operation) was later assigned. This modification worked well in minimizing the mistakes made by inexperienced staff and created a collateral benefit of collecting detailed information about the addresses for further research and analysis.

One way to improve the list of addresses eligible for FAV is to improve the identification and removal of commercial addresses from the AREX address files. The product used was the American Business Information (ABI), Inc. database file of commercial addresses (more than 10 million) based on national telephone directories (both yellow and white pages). Budgetary constraints precluded purchase of the ABI residential file. The use of both files (commercial and residential) would have improved the accuracy of commercial address identification and reduced the size of the FAV eligible address list as well.

It is difficult to gauge the impact of the FAV because the actual review was carried out on a small sample and because of the classification error associated with the imputation based on the regression equation. But there are some things that can be learned from the FAV sample.

The sample addresses were drawn from a list that did not match any address in the initial DMAF match. About 25 percent of the sample addresses found to be valid upon field review were found to be valid as listed. They represent about 10 percent of all FAV eligible addresses. The remaining 75 percent of valid sample addresses (30 percent of all eligible addresses) were found to be valid after lister corrections. It turned out that none of this group matched a Census 2000 address in the second match to the DMAF nor, of course, did the uncorrected valid group. It is not known whether any of these addresses truly represent addresses not in the DMAF or are unmatched as a result of inaccuracies in the address matching process.

Table 16 shows Bottom-up "best" administrative records addresses by FAV status.

Table 16. Bottom-up "best" address by FAV Status.

	FAV Valid	Imputed Valid	Not FAV Eligible	Valid in Second DMAF Match, Only	Total
AREX Occupied	1,084	24,703	855,946	3,775	885,508
	43%	47%	86%	33%	83%
AREX Vacant	1,420	28,242	142,253	7,608	179,523
	57%	53%	14%	67%	17%
Total	2,504	52,945	998,199	11,383	1,065,031
	100%	100%	100%	100%	100%

It is interesting to note that FAV sample addresses selected as best addresses were much more likely to be vacant than addresses that were not FAV eligible. The FAV imputed addresses had occupancy rates that were similar to the sample. The number of persons counted at FAV sample addresses was 2,162 and at FAV imputed addresses, 44,912 for a total of 47,074. Inflating the FAV sample persons by the reciprocal of the average selection probability (i.e. 2,162(1/. 0433)) yields 49,931, much of the difference presumably due to address misclassification as a result of the imputation. However, the closeness of the numbers suggests that a 100 percent FAV would have yielded results similar to the combined sample and imputation scheme.

Including AREX vacant housing in the Census Pull

The AREX address selection rules resulted in almost 180,000 vacant addresses thought to be valid for the AREX test sites. Such addresses that are actually found in the AREX sites through a match to the Census 2000 HDF would appear to be conceptually similar to the addresses included in the Census Pull. Both kinds of addresses represent housing units in the AREX sites for which no administrative records persons were found to be resident. In both cases, it might have been that some addresses were truly vacant on census day and others truly occupied. For the latter, deficiencies in the administrative records or administrative records processing resulted in the persons not being counted or counted at the wrong address.

A match of the AREX vacant addresses to the Census 2000 HDF, in fact, found about 76,000 matched addresses, and almost 67,000 were occupied in Census 2000. (Refer to the analysis in Section 3 of this report.) Of course, some of these persons may be the same as persons counted at other administrative records addresses; but the same could be said for the persons found at the Census Pull addresses. Therefore, in a Bottom-up process, both types of addresses should have been canvassed; and the AREX vacant addresses that matched addresses in the Census 2000 HDF should have been included in the Census Pull.

Unduplication after the Census Pull

There should have been an unduplication of individuals after the Census Pull by matching persons obtained in the Pull with those from the administrative records lists. The presence of duplicate individuals is suggested not only by the overcounts of adults shown in various tables, but also by a comparison of the total number of Bottom-up addresses with the number of Census 2000 addresses in the test sites. The total number of addresses in the Bottom-up was 1,217,810 -- 1,065,031 administrative record addresses and 152,779 from the Census Pull. The number of Census 2000 address in the test sites was 1,094,204. Thus, there were 123,606 more addresses in the Bottom-up enumeration than in Census 2000.

One way to accomplish the unduplication would be to search and verify the SSNs for the individuals in the Census Pull and compare them with the SSNs of the individuals in the administrative records lists. This might not be completely effective because being part of the Census Pull suggests that blocking on address will not facilitate the SSN search. Alternatively, the Census Pull individuals could be matched directly with the administrative record list blocking variously on such variables as surname and date of birth.

When duplicate individuals were found, the Census Pull could be taken as more accurate and the individuals would be removed from the administrative records address. This approach could result in some additional vacant addresses, so that the process might have to be repeated several times in order to identify the "best" address for all persons. In the end, there could be vacant administrative record addresses that should have been filled by persons erroneously located outside of the AREX sites in the administrative records systems. This would imply that a national unduplication would be part of a full Bottom-up census. Such an unduplication was out of scope for the experiment.

3. AREX OUTCOMES AND HOUSEHOLD EVALUATIONS

3.1 Introduction

The evaluation of the numerical findings of the AREX was twofold. First there was a comparison of the results of the Top-down and Bottom-up enumerations with the Census 2000 enumeration in the experimental test sites (Heimovitz 2002). This analysis progressed from large geographic areas to small geographic areas, beginning with the five test site counties and ending with Census 2000 blocks within the sites. The outcomes evaluation tried to disentangle the influence of demographic change and AREX processing, coverage and data quality issues, while presenting basic enumeration statistics. Below the county level, the analysis focused on the Bottom-up enumeration because the county-level analysis was sufficient to show the evident weaknesses of the Top-down process. Section 3.2 provides some of the highlights of portions of Heimovitz's report; there was also a regression analysis that is omitted here.

The primary goal of the second evaluation was to assess the accuracy of households assembled from administrative records by comparing them to Census 2000 enumeration results at the same addresses (Judson and Bauder, 2002). This was a particularly important analysis for the type of design that the AREX mounted because the completion of an administrative records enumeration by canvassing addresses not found in the records provides little opportunity to correct enumeration errors in the administrative records themselves. Thus, it was important to learn as much as possible about the strengths and weaknesses in the administrative records households with an eye toward future improvements.

In the course of the household-level analysis, some preliminary information about a possible use of administrative records in a conventional census was obtained. The question of interest was: Under what conditions can administrative records households be substituted for conventional Nonresponse Followup (NRFU) households, or households for which occupancy status and household demographics were wholly imputed ("unclassified" households)? This assessment was carried out by matching the demographic composition of AREX households to Census 2000 households which were difficult to enumerate in Census 2000. In addition to a descriptive analysis, there was a prediction-based approach to assess the ability to predict when an AREX household is likely to demographically match a census household. Section 3.3 provides a summary of the results in the report by Judson and Bauder.

3.2 AREX enumeration outcomes

Methodology

Concept

The enumeration outcomes analysis provides measures of how well AREX replicates Census 2000 results at county and subcounty levels focusing on key demographic characteristics that are important for decennial census requirements but also relate to the possible use of administrative records for intercensal and small-area estimation. A series of research questions provides a conceptual outline of the basic elements of the evaluation. General questions at larger geographies are posed first:

- How well does AREX measure total census population at the county level, and how do the results differ by whether the Top-down or Bottom-up approaches were used?
- How do county-level differences between AREX and census differ by age, race, sex, and Hispanic origin, as well as between the Top-down or Bottom-up approaches?

A related question, how well does AREX measure the voting age population (age 18+) of state legislative districts, is discussed in Heimovitz, 2002.

In a decennial census, total population counts are needed for congressional apportionment. The voting age (18+) population by race and Hispanic origin potentially meets the data requirements for legislative redistricting. Population counts of persons under age 18 are needed by states for planning purposes and estimating child poverty rates. Greater differences between AREX and census counts are more likely at smaller geographies. But focusing on smaller geographies allows more detailed analyses of neighborhood characteristics and whether these attributes are linked with AREX-Census 2000 differences: In particular, how does the accuracy of tract and block counts compare to county results?

Outcome measures

The terms ‘undercount’ and ‘overcount’ describe how well AREX counts match Census 2000 results and have no further connotation. That is, undercounts and overcounts reflect any of several problems, including coverage issues, coding, and processing errors. Outcome and predictor constructs are distinguished and used to highlight AREX-Census 2000 Bottom-up and Top-down differences. The outcome measures used in this consolidated report are limited to the simple count differences between AREX and Census 2000 counts and to the algebraic percent error (ALPE). The full outcomes analysis (Heimovitz 2002) provides additional measures.

Difference

The simple difference between AREX and Census 2000 gauges the county-level over and under-counts:

$$\text{DIFF}(A_i,C_i) = A_i - C_i$$

where:

A_i = AREX tallies in county
C_i = Decennial census tallies in county

Algebraic percent error (ALPE)

AREX and Census 2000 counts are the inputs for calculating the algebraic percent error for the i^{th} county, tract, or block:

$$\text{ALPE}(A_i, C_i) = \frac{A_i - C_i}{C_i}$$

Where:

A_i = AREX tallies in the i^{th} county, tract, or block; and
C_i = Decennial census tallies in the i^{th} county, tract, or block

Two problems can occur when computing ALPEs: zero blocks and inflated ALPEs. Zero blocks occur when AREX reports in a particular block at least one person having a particular characteristic but census does not. Because Census 2000 is being used as the standard and is the denominator, ALPEs for zero blocks are undefined. For the purpose of block comparisons, zero blocks are omitted from the analyses. However, county and tract-level counts and comparisons include these blocks because they are aggregated at larger geographies.

Inflated ALPEs can sometimes occur when Census 2000 blocks have very small counts and tend to produce large, positive ALPEs, despite small differences between AREX and Census 2000 counts. For example, C=1 and A=3 yields an ALPE=2. Such a large ALPE is quite unlikely when the size of C--the number of persons enumerated in the census area--is large. Small census counts are not unlikely, for example, for racial minorities in sparsely populated areas. To reduce the impact of unusually large ALPEs, ALPEs were trimmed (topcoded) by setting all values greater than the 95th percentile of the ALPEs across the areas in the analysis to the value of the 95th percentile. Still, care should be taken in interpreting results for those analyses where the population is sparsely populated within the geographic units of interest.

There is an additional problem when computing differences or ALPEs for racial subpopulations. The problem stems from the differences between AREX and Census 2000 classifications. Both AREX and Census 2000 have the four traditional categories: White, Black, American Indian/Alaskan Native, and Asian/Pacific Islander[8]. But Census 2000 permits respondents two additional options: "multiple race," and "Other race." These additional groups were quite small for Maryland. For Colorado, the multi and other race groups were much larger, encompassing more than 8 percent of the Census 2000 population for El Paso County. In the following outcomes analysis, no attempt was made to distribute either additional race category across the four common categories. Excluding Census 2000 respondents with multi or other race could result in positive differences and ALPEs for race subgroups, especially for minority groups, that might not have occurred had the AREX and census classifications been the same.

Descriptive analyses

This section is intended to be a top-level, descriptive summary of AREX-Census 2000 differences, by county, tract, and block. County-level counts and proportions are compared and display the raw, untransformed numbers not shown in the multivariate analyses. The count differences describe the aggregate under- and over-counts of age, race, sex, and Hispanic origin categories, while the ALPEs show the contribution these categories have on the under- and overcounts. One important aspect of the bivariate analyses is the ecological variation within the AREX counties. Thematic maps profile the heterogeneous AREX-Census 2000 differences in block-level total population counts.

AREX Top-down counts include persons later identified in Bottom-up as group quarters residents; Bottom-up and Census 2000 counts exclude group quarters residents and differ somewhat from counts in earlier tables for which there were no exclusions.

<u>County-level count results</u>

Total population

Total population results for the two Maryland counties and three Colorado counties are reported in Table 17.

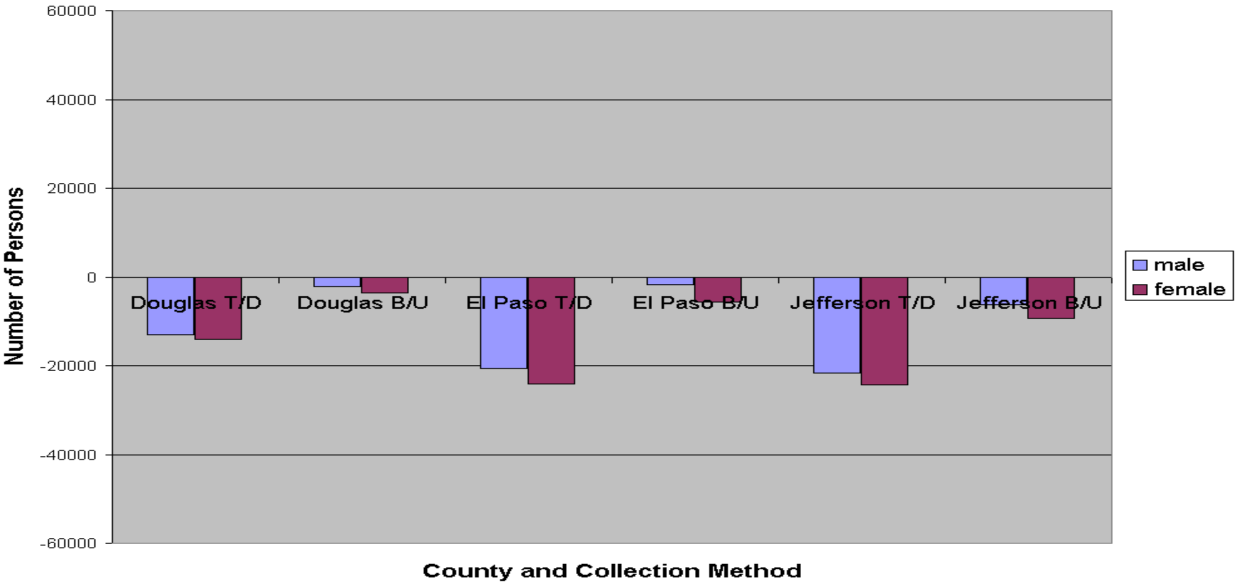
[8] Multiple Census 2000 categories were combined for Asian/Pacific Islander.

Table 17. Top-down and Bottom-up Counts of Total Household Population by County

	Top-down Results				Bottom-up Results			
	AREX	Census	Difference	ALPE	AREX	Census	Difference	ALPE
Baltimore County	696,183	736,652	-40,469	-5.5%	728,205	736,652	-8,447	-1.1%
Baltimore City	570,648	625,401	-54,753	-8.8%	636,729	625,401	+11,328	+1.8%
Douglas County	148,270	175,300	-27,030	-15.4%	169,640	175,300	-5,660	-3.2%
El Paso County	456,891	501,533	-44,642	-8.9%	494,253	501,533	-7,280	-1.5%
Jefferson County	473,495	519,326	-45,831	-8.8%	503,622	519,326	-15,704	-3.0%

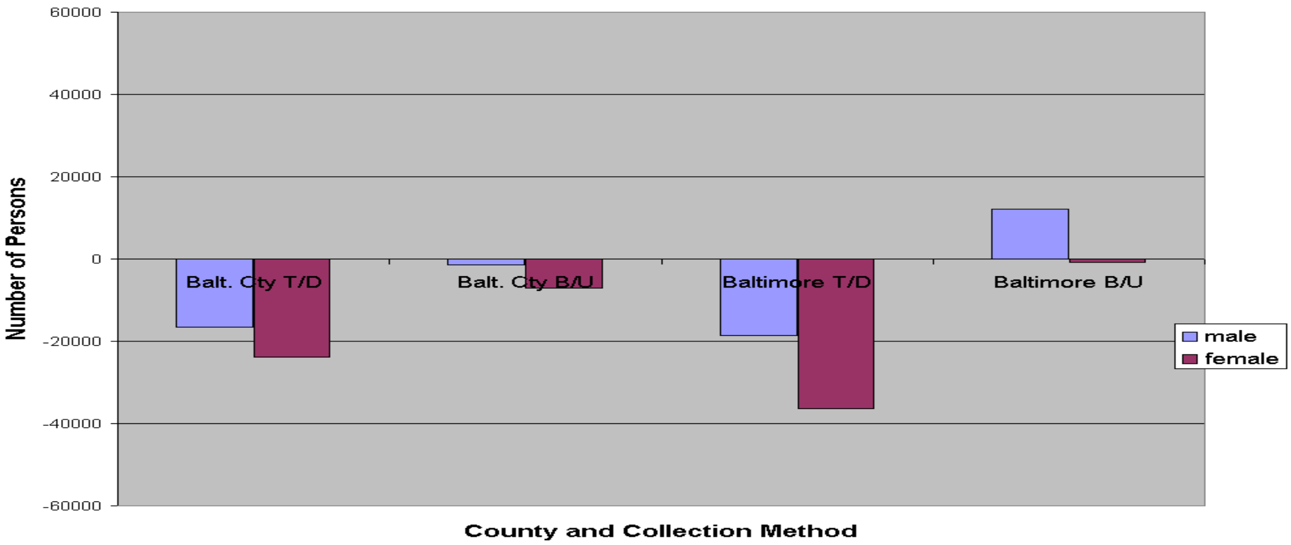
AREX undercounted all five counties in the Top-down and four of five counties in Bottom-up. The greatest Top-down differences were in Baltimore City and Jefferson County. Bottom-up undercounts are much smaller than Top-down undercounts in all five counties for total population and demographic characteristics.

Figure 2. Net Population Difference by Sex, County, and Collection Method--CO.



Sex

Figure 3. Net Population Difference by Sex, County, and Collection Method--MD.



Males and females are undercounted by the Top-down method in all five counties. Bottom-up undercounts are much smaller for all counties, and males are overcounted in Baltimore City. (Baltimore CTY is Baltimore County.)

Age

Figure 4. Net Population Difference by Age, County and Collection Method--MD.

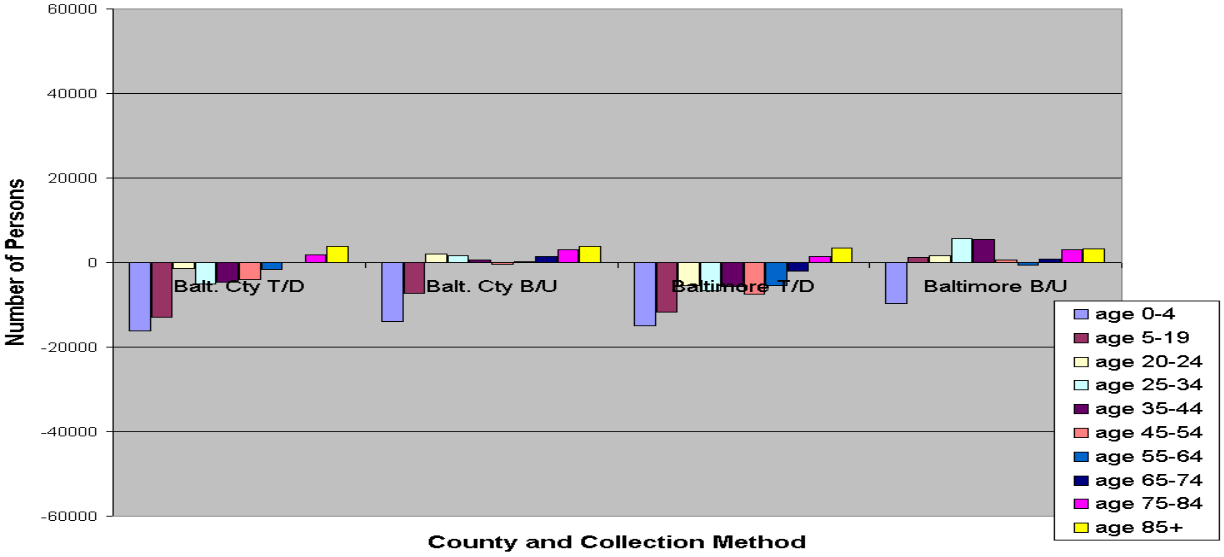
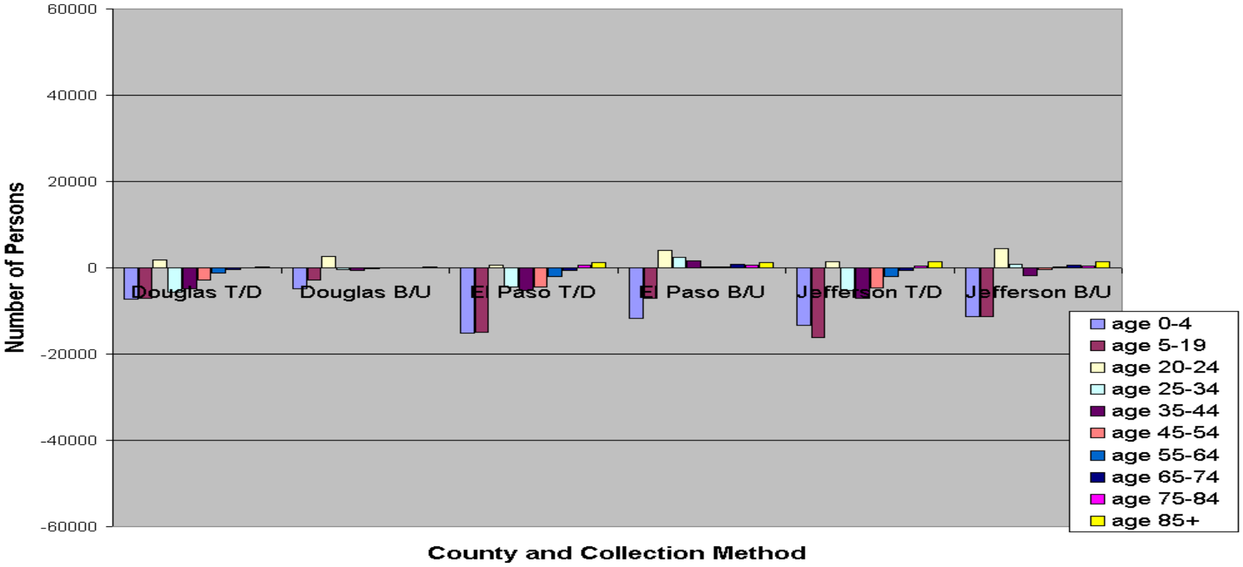


Figure 5. Net Population Difference by Age, County and Collection Method--CO.



In the Maryland counties, Top-down overcounts the 75+ population and undercounts other age groups; Bottom-up overcounts the 20-44, and 65+ age groups and undercounts all other age groups. In both Maryland and Colorado, Top-down undercounts are greatest for the 0-19 age groups and show the greatest improvements for Bottom-up counts relative to Top-down. In the Colorado counties, generally, age 20-24 and 65+ age groups are overcounted and other age groups are undercounted for both Top-down and Bottom-up methods.

Race

In the Maryland counties, Hispanics were overcounted and other minority race groups were generally undercounted in Top-down and Bottom-up. In the Bottom-up method, Whites and Blacks were overcounted in Baltimore City where Blacks are a majority. In the Colorado counties, Blacks and APIs were generally overcounted while other race categories and Hispanics were undercounted in Top-down and Bottom-up methods.

Figure 6. Net Population Difference by Race, County and Collection Method--MD.

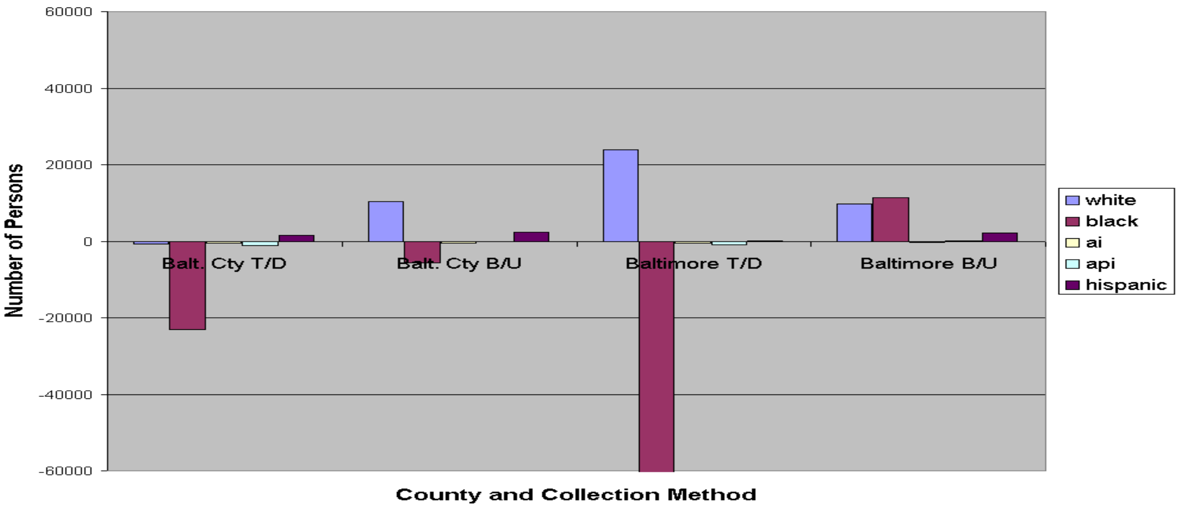
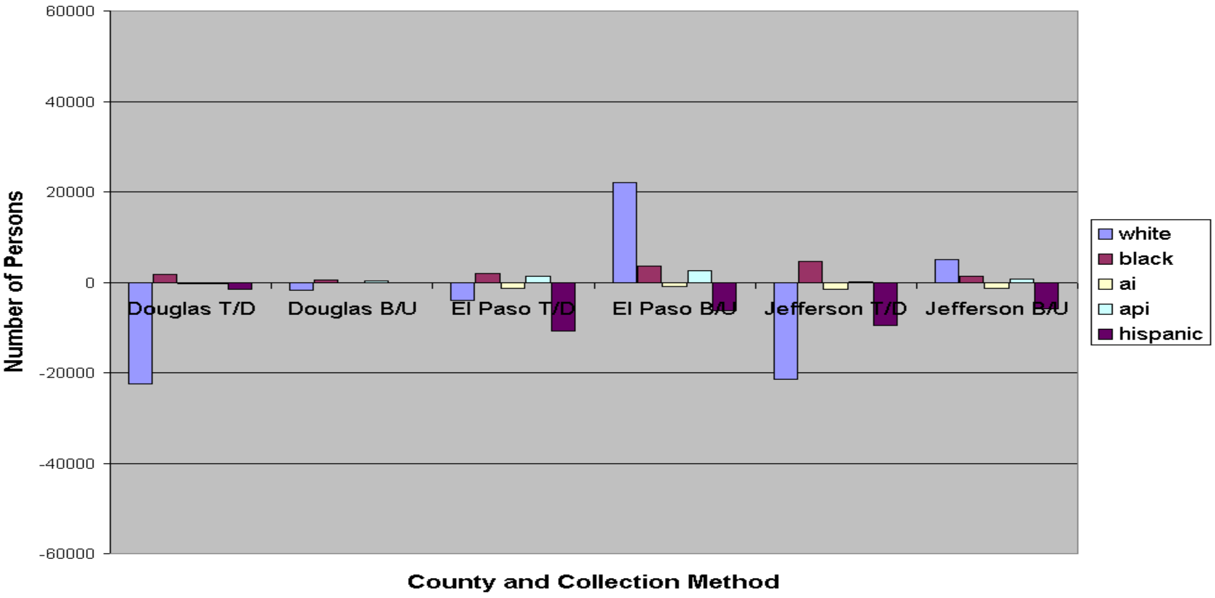


Figure 7. Net Population Difference by Race, County and Collection Method--CO.



One general pattern from tables and figures above is the relationship between population shares and AREX under- and overcount. Race and Hispanic origin groups with smaller shares tend to be overcounted, and groups with larger shares tend to be undercounted. Examples of overcounts are Hispanics in the Maryland counties, Whites in Baltimore City, and Blacks and APIs in the CO counties.

The very large Top-down undercount of Blacks in Baltimore City is due largely to the inappropriate use of the race model for children in the Top-down process. The Black count changes dramatically in the Bottom-up in which children with unknown race are generally assigned the race of the primary taxpayer.

<u>County-level ALPE results</u>

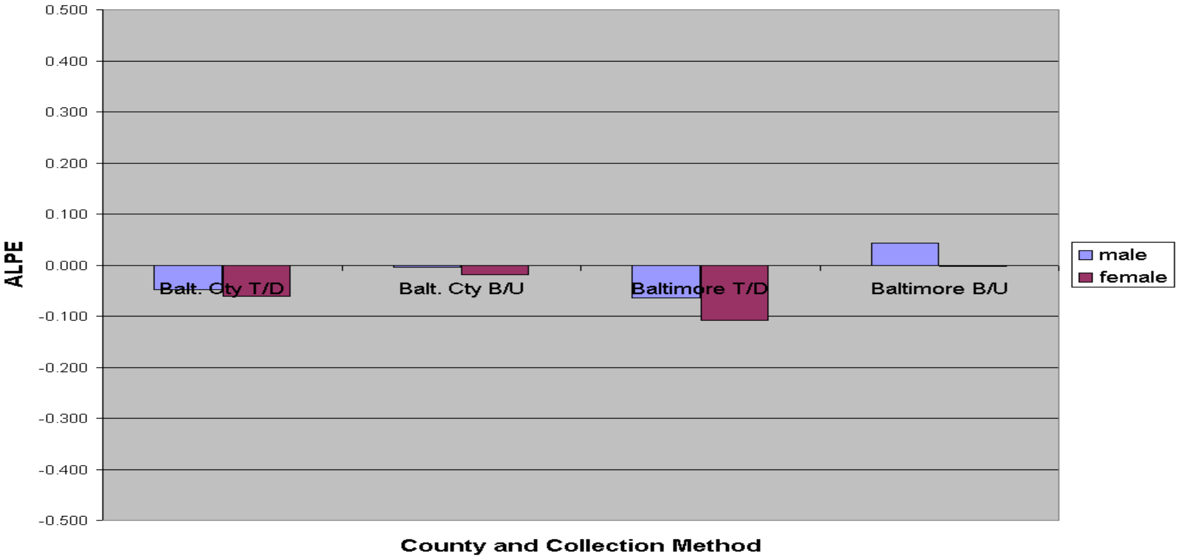
The county-level analysis builds on the AREX-Census 2000 count results by examining the algebraic percent error (ALPE). The ALPE measure provides a different view of the county-level results because the calculation method uses census group totals as bases and provides a standardized gauge for comparing differences between Top-down and Bottom-up, as well as between counties.

Total Population

All county Bottom-up ALPEs were smaller than Top-down ALPEs; Bottom-up ALPE improvements were variable: both Douglas County and Baltimore City had Top-down ALPEs of -8.8 percent, but Bottom-up for Douglas County was -3.2 percent compared to +1.8 percent for Baltimore City. The smallest total population Bottom-up ALPE was in Baltimore County (-1.1 percent); the largest Bottom-up ALPE was in Douglas County (-3.2 percent).

Sex

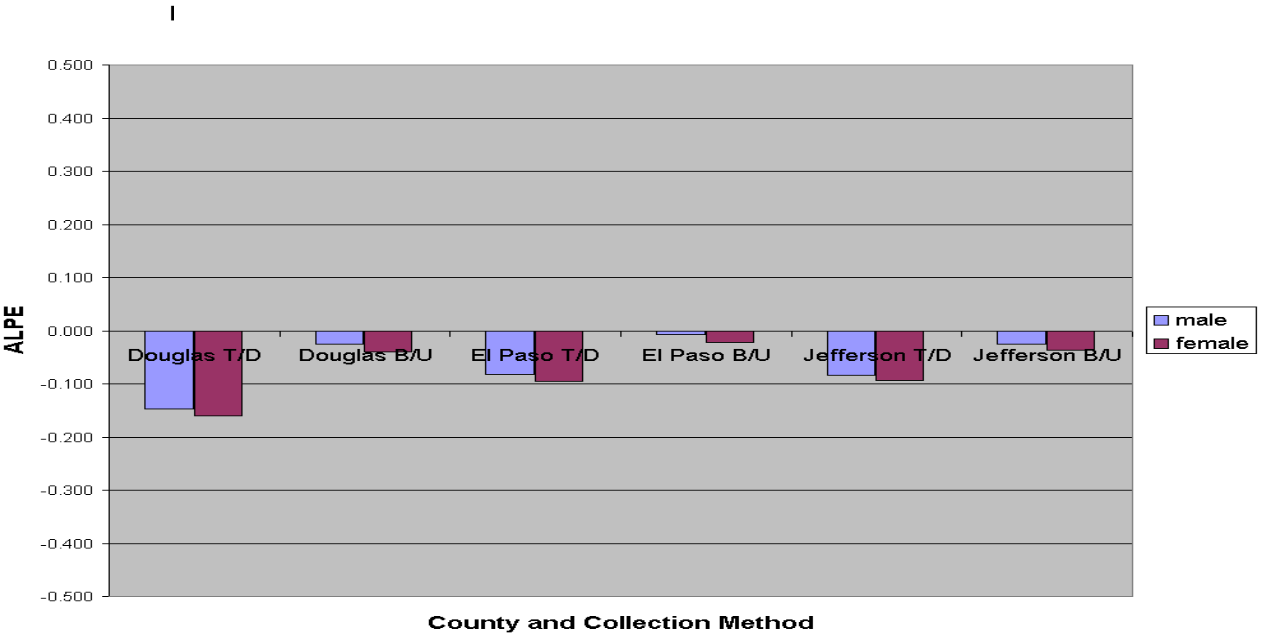
Figure 8. Sex ALPE by County and Collection Method--MD.



Male and female Bottom-up ALPEs were relatively small in all five counties and ranged from – 4.8 to + 4.2 percent.

Sex proportions were undercounted in all counties (except Baltimore City males) and generally are unbiased, reflecting the magnitude of total county-level proportions. Female undercounts were slightly worse than male undercounts and generally had a marginal difference of less than 2 percent in Bottom-up. Some women may be less active within the administrative records systems. For example, some studies indicate that lifetime participation in the labor force varies by a woman's child raising and care giving experiences, health status, and race/ethnicity (Flippen and Tienda, 2000). However, lower mortality rates for women might offset lower labor force participation with respect to AREX/Census 2000 comparisons.

Figure 9. Sex ALPE by County and Collection Method--CO.



Age

Generally, younger age groups (especially the 0-4 age group) had the largest negative ALPEs in all five counties. Bottom-up ALPEs for the 0-4 age group ranged from –33.9 percent in Jefferson County to –23.4 percent in Baltimore City. Older age groups (65-74, 75-84, and 85+) tended to have positive ALPEs that increased by age. Bottom-up ALPEs were generally smaller due to the Census-pull households that replaced unmatched Census 2000 addresses.

Figure 10. Age ALPE by County and Collection Method--MD.

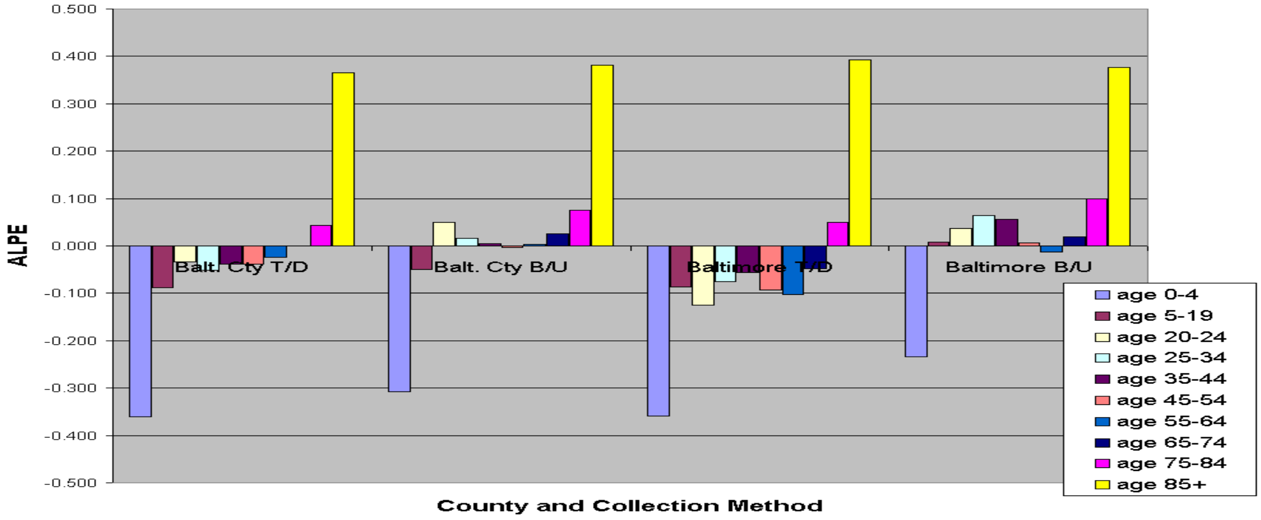
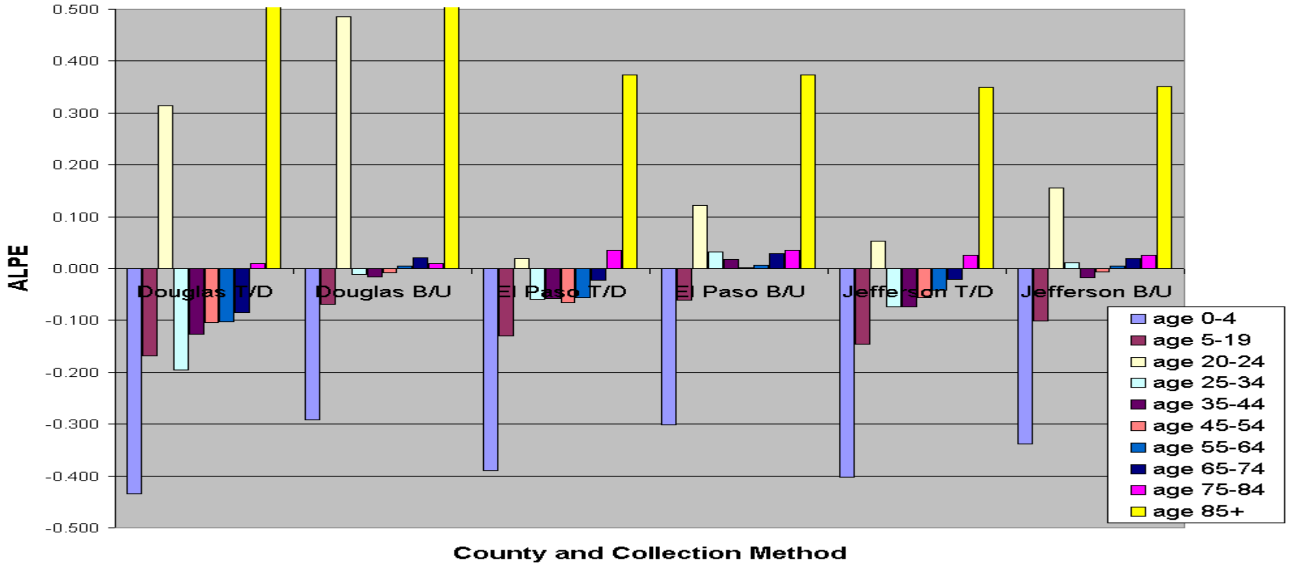


Figure 11. Age ALPE by County and Collection Method--CO.



The large negative ALPEs for children and the large positive ALPES for older groups are due mostly to the weaknesses in the administrative records discussed in Section 2: missing births and deaths and migration as a result of the cutoff dates of the administrative record files used in the AREX, missing dependents on IRS 1040s, and missing children of parents who did not have to file 1040s or were otherwise not found in the administrative records. Persons aged 65+ were generally overcounted in all five counties, and persons age 85+ displayed Bottom-up overcounts ranging from about 2 percent to 36 percent--77 percent in less-populated Douglas County. Because the 85+ population is relatively small, the denominators of the ALPE calculations are likely to be small and potentially inflate ALPE measures.

The 20-24 year age group also has large positive ALPEs in some of the AREX counties. This might be due to the handling of special populations to which this age group belongs: college and university population, and the military. College-age persons whose residence may have been reported at a parent's IRS tax address may actually reside on a campus in a different area. Removing group quarters from the Census 2000 counts but not from the Top-down counts would bias Top-down ALPEs in the positive direction. Removing group quarters from the Bottom-up counts would still leave dependents claimed on IRS 1040 at the wrong location with respect to decennial residency rules.

Douglas County appears to be a special case. The Census 2000 population age 20-24 is 3.1 percent, less than half that of Colorado (7.1 percent) and the national average (6.7 percent). But the Air Force Academy and several other schools are located in Douglas County. The large Top-down ALPE may be due to the fact that group quarters were not removed from the administrative records. Although there was an attempt to remove group quarters from Bottom-up enumeration, the large Bottom-up ALPE for age 20-24 suggests this may not have been fully successful.

Race

It is difficult to interpret Top-down race ALPEs because of the confounding effects of general undercounts, especially for children, and the use of the race model for children under 15 with "other" or unknown race in the administrative records. The following discussion will focus on the Bottom-up results.

There are a number of reasons for the patterns of Bottom-up race and Hispanic origin ALPEs. First is the use of the race model. As discussed earlier, the race model was a national-level model and variation about its predictions can be expected. The use of the model in small geographic areas would tend to overstate the number of persons in those race groups that are less than the national average and understate the number of persons in groups that are above the national average. When modeled race was assigned to children from an adult in the same household, the result would be reinforced.

Figure 12. Race ALPE by County and Collection Method--MD.

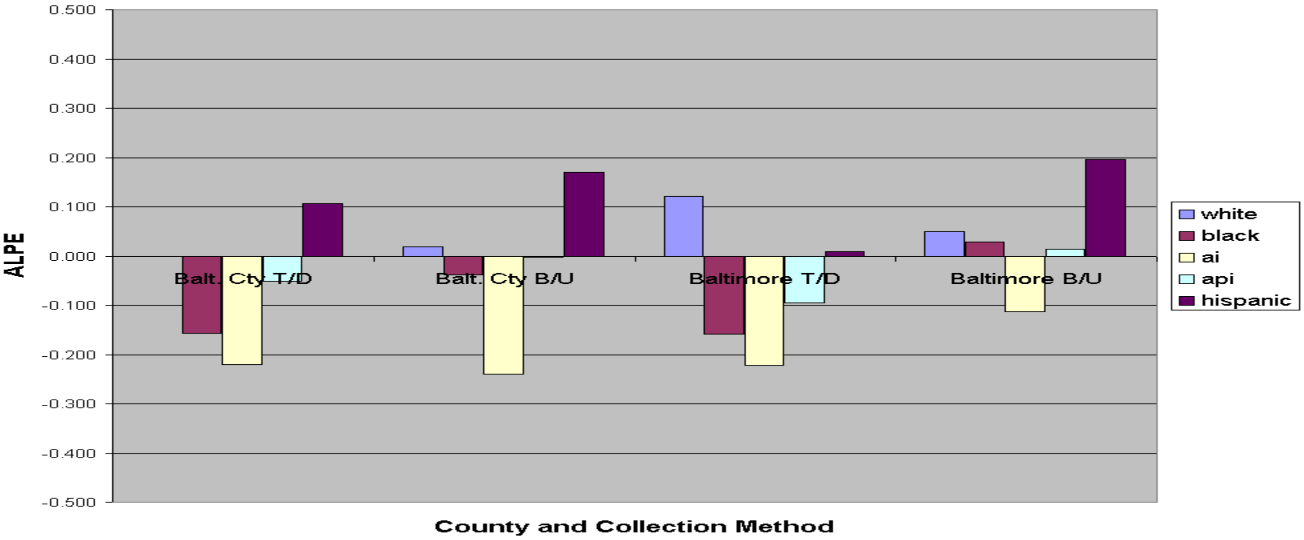
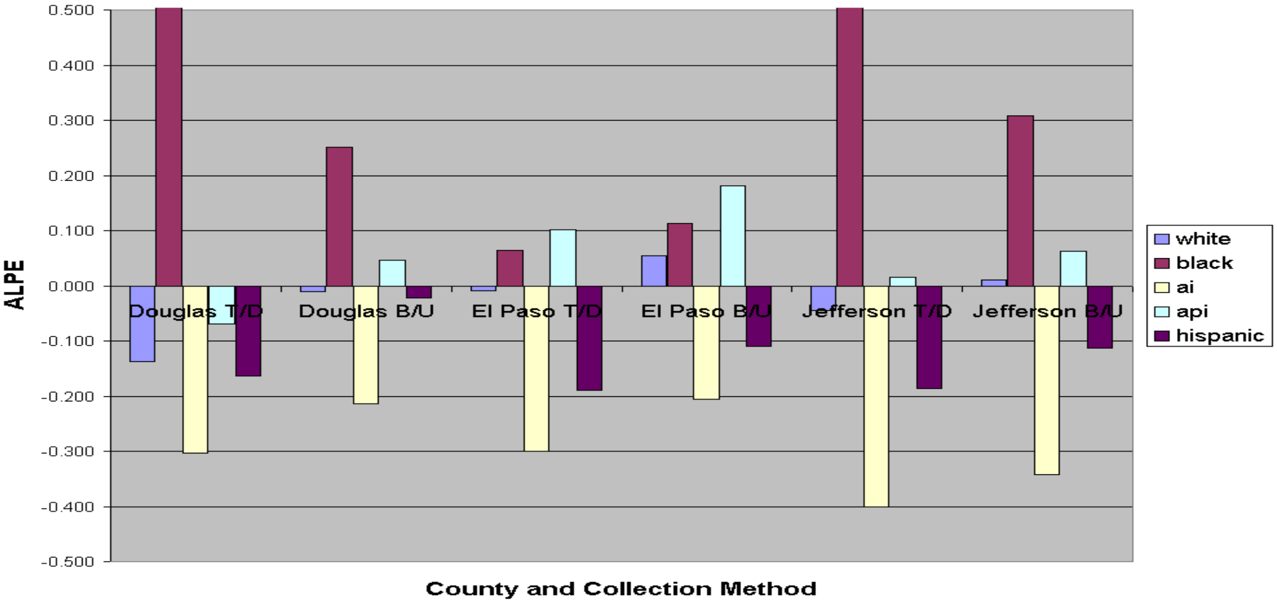


Figure 13. Race ALPE by County and Collection Method--CO.



It is important to keep in mind that the race model was used for only those adults whose administrative records did not provide a race other than “other” or unknown. Table 3 in Section 2.2 shows the proportion of adults and children with imputed race in each of the test sites. The number of adults with imputed race ranged from about 3 to 9 percent and was substantially lower in Maryland than Colorado. For Hispanic origin, the imputation was used well over 90 percent of the time because, for the most part, administrative records provide no direct measure of ethnicity.

Other factors possibly affecting Bottom-up ALPE race patterns were: The possible correlation between weaknesses in AREX population coverage and race or Hispanic origin, unaccounted for migration and demographic changes due to the age of the administrative records files, the possible duplication of persons due to the Census Pull, the problem of comparing AREX and Census 2000 race groups because the latter allows "multi" and "other" and the former does not, and the positive ALPE bias for cells with small denominators. Examining several of the race ALPE results shows the complexity of the possible explanations.

For the Bottom-up, Black ALPEs were positive in all three Colorado counties and Baltimore City and negative in Baltimore County (where blacks are a large minority race group). The overcount of Blacks in Colorado was most likely due to the race model because the proportion of Blacks in Colorado was much smaller than the national average; and at the same time, the proportion of adults in Colorado with imputed race was relatively high, ranging from 6 to 9 percent. The undercount of Blacks in Baltimore County might be due in part to the use of the race model; but it might also be due in part to the migration of Blacks from Baltimore City to the County in the period between the administrative records cutoffs and April 1, 2000. The reasons for the overcount of Blacks in Baltimore City are less clear but might also be due to unaccounted for migration of Blacks from the city. An overcount is the reverse of what would be expected if the race model were the main cause, and the proportion of adults with modeled race was under 3 percent.

The ALPEs for APIs were positive in all three Colorado counties and Baltimore City and negative in Baltimore County. This would appear again to be a race model effect, except for Baltimore County, because nationally, all five counties have API proportions below the national average. Evidently, the net effect of these differences increased the size of the Census 2000 API counts enough so that API ALPEs for all of the AREX sites would have been negative had they been calculated from these distributions. In any case, it is simply a matter that ALPE is sensitive for small population subgroups.

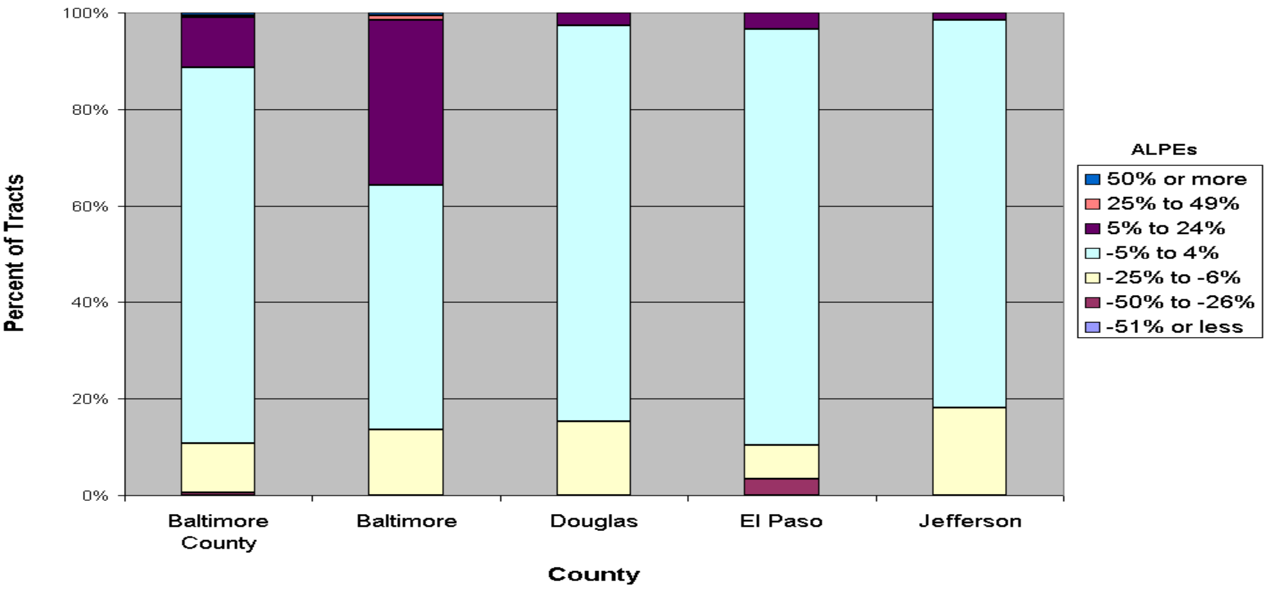
Concerning APIs, the substantial negative ALPEs in all counties were not unexpected. Identifying AIAN race is weak in the administrative records, except, in areas around reservations, and AIAN prediction was the weakest part of the race model as well.

Hispanic ALPEs were positive in both MD counties where they are a small minority group and negative in all three CO counties where Hispanics are the largest minority group. The model for Hispanic origin was applied to about 97 percent of adults in Maryland and is most likely the reasons for the substantial overcounts there. Again, one might have also expected small overcounts in Colorado were model use the main factor. (See the discussion of Hispanics in Section 2) But the substantial undercounts suggest that other factors may be at work such as high birth rates and net in-migration of Hispanics in the period missed by the administrative records used in the AREX.

Tract ALPE distributions

Error! Reference source not found. shows the ALPE distributions for the five AREX counties. In all sites other than Baltimore City, more than 70 percent of tracts had AREX total population counts within +/-5 percent of census results, and more than 95 percent of tracts had counts within 25 percent of census results. Baltimore City had less accurate results with about 50 percent of tracts exceeding +/-5 percent of census results. A larger proportion of tracts had moderate and large ALPE undercounts (less than –5 percent) compared to overcounts.

Figure 14. Distribution of Tracts with Under- and Overcounts of Total Population.

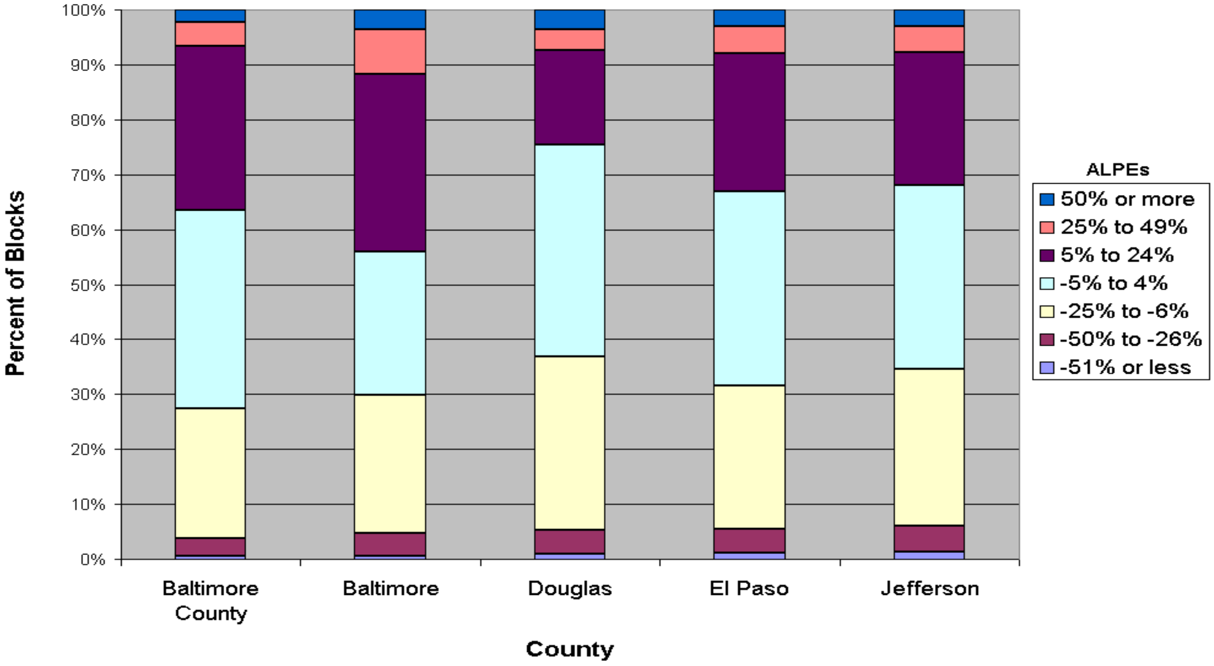


Though the tract-level ALPEs for the total population resemble county-level results, the distributions indicate more Baltimore City tracts were overcounted. It is unclear whether these overcounts are related to persons who were actually uncounted in the census, or more likely, weaknesses in AREX processing. Households may have been added through the Census Pull process that replaced unmatched addresses that existed in other tracts or addresses.

Block ALPE distributions

The block-level ALPE results describe the accuracy of counts at the smallest geographic level and relative to counties and tracts. The main problem with this type of comparison is the ALPE denominator potentially inflates block-level ALPEs for small population subgroups and especially minorities. This inflation is likely to be greater than found in the tract-county comparisons. A second issue affecting comparisons is the exclusion of blocks where Census 2000 did not identify persons with a particular attribute (zero blocks). County and tract ALPEs include blocks with zero counts because these blocks were accumulated into larger geographies. However, the block-level ALPEs used the reduced set of blocks and the results may be quite different when comparing the ALPEs at various geographies.

Figure 15. Distribution of Blocks with Under- and Overcounts of Total Population.



AREX was less accurate in estimating blocks than tracts in all counties. Population totals for 18 to 39 percent of blocks were within 5 percent of Census 2000, and about 85 percent were within 25 percent of the census. Douglas County had the best results at the 5 percent criterion and Baltimore County was best at the 25 percent criterion. In the Maryland counties, slightly more blocks had moderate or large overcounts (ALPEs exceeding 5 percent), compared to the Colorado counties where more blocks had moderate undercounts (-5 to -24 percent).

The AREX counts were less accurate at the block-level. Population counts are likely to be less accurate in smaller areas due to incorrect assignment of households at tracts and blocks that average out for county-level counts. This is demonstrated by the greater number of moderate and large ALPEs and indicates how smaller denominators and AREX processing weaknesses influenced the comparisons. Though zero blocks were excluded and fewer blocks met the 5 percent criterion, a large proportion of blocks met the 25 percent criterion in all five counties.

Geospatial Tract-Level Heterogeneity

Error! Reference source not found. and **Error! Reference source not found.** exhibit the geographic distribution of AREX-Census 2000 tract ALPEs for total population counts. Baltimore City is the nucleus of the MD AREX site and the most urban of all the sites. It has numerous tracts with large under and overcounts. The tract-level total population was clearly measured better in Baltimore County. There is also evidence of tracts clustering by size of under and overcounts. Downtown Baltimore and Towson include islands of moderate and large undercounts, while clustered moderate overcounts are more frequent in other parts of the City and County. Denver, in the north, and Colorado Springs are metropolitan centers in the CO site (Map 2). Generally, the tract-level CO population was counted more accurately in the suburbs of each city, while urban and rural tracts tended to have moderate undercounts.

Figure 16. AREX - Census Total Population ALPEs: Maryland Tracts.

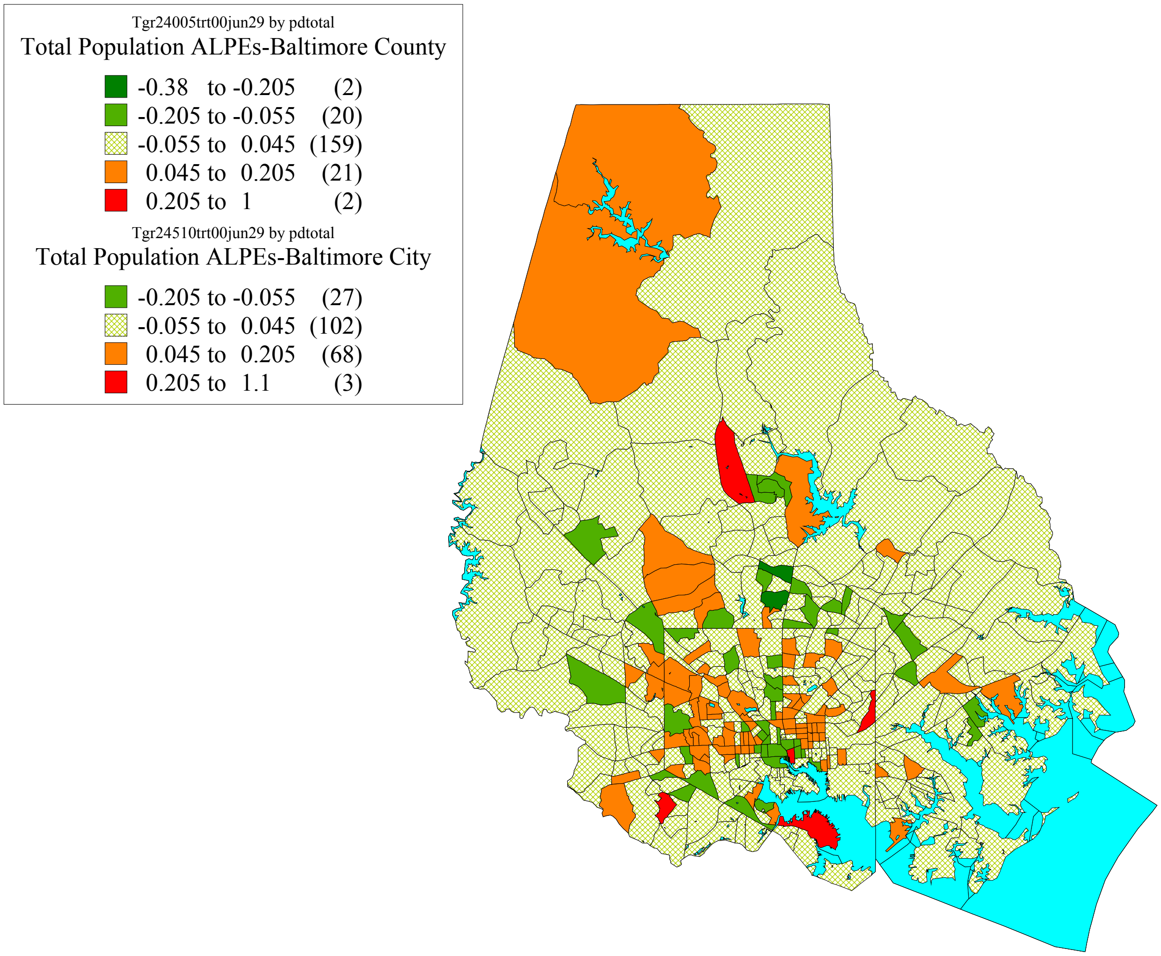
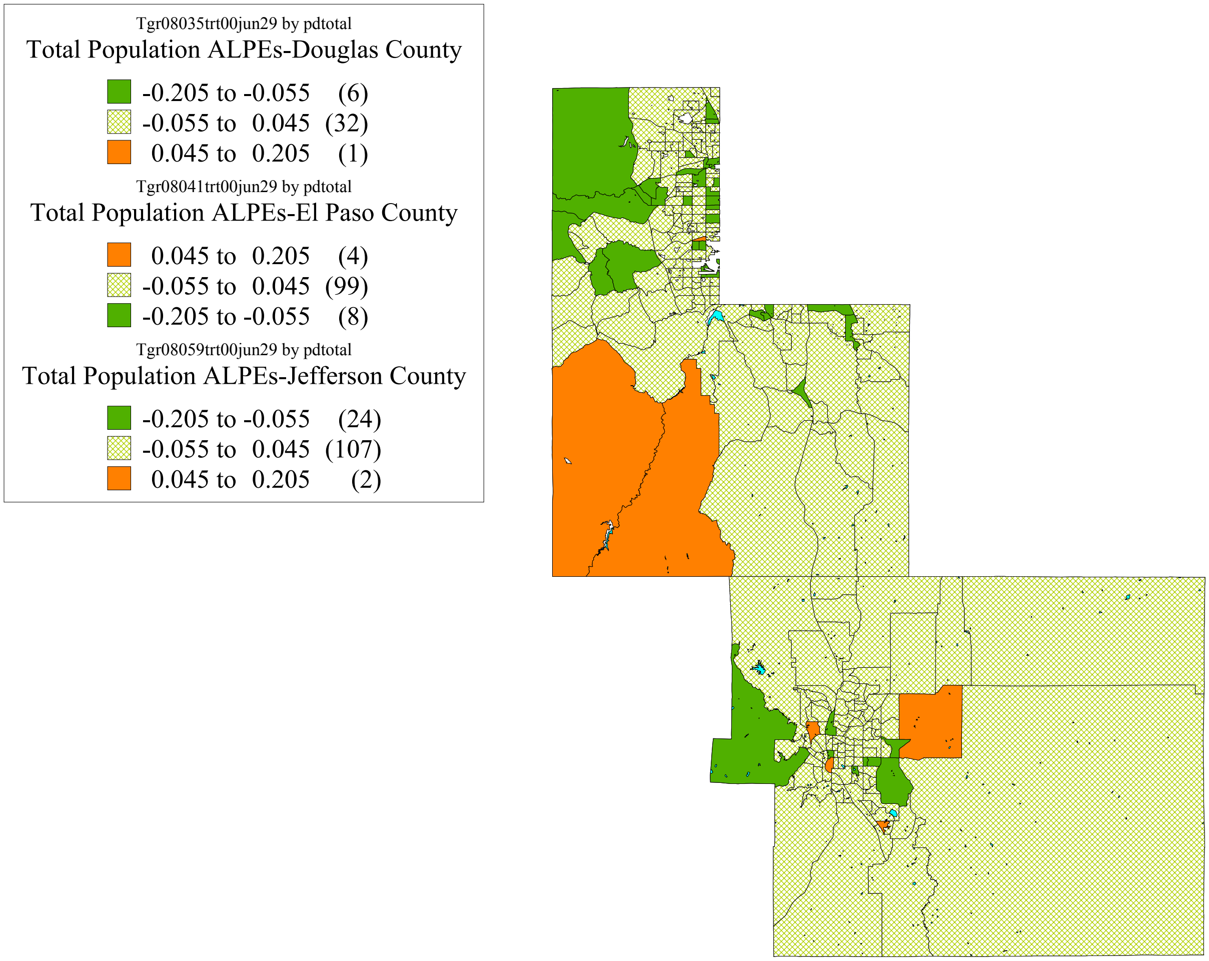


Figure 17. AREX - Census ALPEs for the Total Population: Colorado Tracts.


Tgr08035trt00jun29 by pdtotal
Total Population ALPEs-Douglas County
🟩 -0.205 to -0.055 (6)
▨ -0.055 to 0.045 (32)
🟧 0.045 to 0.205 (1)
Tgr08041trt00jun29 by pdtotal
Total Population ALPEs-El Paso County
🟧 0.045 to 0.205 (4)
▨ -0.055 to 0.045 (99)
🟩 -0.205 to -0.055 (8)
Tgr08059trt00jun29 by pdtotal
Total Population ALPEs-Jefferson County
🟩 -0.205 to -0.055 (24)
▨ -0.055 to 0.045 (107)
🟧 0.045 to 0.205 (2)

Summary and conclusions

The forgoing analysis provided measures of how well AREX replicated Census 2000 results at several geographic levels focusing on key demographic characteristics that are important for a decennial census. As expected, the Bottom-up method performed better than the Top-down method because of the simulated canvassing of households (the Census Pull) at addresses not found in the administrative records. The Bottom-up process undercounted total population in all sites except Baltimore City. Algebraic percent errors for county-level population totals were less than 5 percent though the results were not as good for subcounty and demographic subgroups.

If the Bottom-up process were unbiased and counted all demographic groups in the same way, ALPEs for all demographic categories would have had the same relative size. As with the total population, males and females were undercounted in all sites except Baltimore City, but the female undercounts were slightly greater than male undercounts. Age group ALPEs show more variability with most groups undercounted except the 20-24 group and the oldest age groups. Generally the size of the undercounts increased with decreasing age. These patterns did not appear to be site-specific and are the result of the weaknesses of the administrative records and certain AREX processing decisions as discussed in Section 2. Overcounts for the oldest old and undercounts for the youngest persons suggest that much more timely birth and death information must be obtained. And the special enumeration requirements for populations such as college students, the military and persons in nursing homes must be incorporated into administrative records processes.

Bottom-up tract-level total population ALPE results indicated a good correspondence between AREX and Census 2000 (70 percent of tracts met the 5 percent criterion; and 95 percent met the 25 percent criterion), though a sizable number of tracts had moderate and large ALPE undercounts. The block-level ALPE results provided the least accurate measure of total population (38 percent of blocks met the 5 percent criterion; and about 85 percent met the 25 percent criterion), compared to tract and county results[9].

The regression results confirm some of the key findings from the univariate and bivariate analyses. Among the mobility variables, both vacancy rate and rental rate were associated with under and overcounts. Generally, rental rate had a greater association with undercounts and vacancy rates had a greater association with overcounts in both AREX sites. As observed in the bivariate analyses, large proportions of persons under

[9] From data not shown in this report (but available from Heimovitz, 2002) ALPE results for sex and age were similar for tract and county analyses. Baltimore City had the worst results for total and demographic ALPE measures but the most accurate results for blacks. However, Baltimore City also had the largest proportion of census pull records and smallest proportion of imputed black race codes. For the race/Hispanic minority groups, the relative size of the minority population in the tract was associated with how well AREX simulated Census results. Tracts with small minority proportions were more likely to have moderate or large positive ALPEs than other tracts.

age 5 and 20-24 were associated with undercounts in both sites. And in CO, large proportions of persons age 65+ were associated with overcounts, other factors held constant (Heimovitz, 2002).

3.3 Household-level analysis

Methodology

Concept

The general goal of the household-level analysis (Judson and Bauder, 2002) was to assess how well households formed from administrative records matched those from Census 2000 at the same addresses in the Hundred Percent Detail (HDF) file. The analysis did not include group quarters or the households found at addresses not in the administrative records files. An assessment of group quarters was beyond the scope of this analysis because AREX did not mount the operations that would have been needed to enumerate special populations in an administrative records census context. And, the Bottom-up "NRFU" households could not be evaluated because the canvassing was simulated by simply including the Census 2000 households at the relevant addresses.

The household-level analysis assessed the ability of AREX administrative records households to match the demographic composition of all households, but there was a special focus on Census 2000 households that required a nonresponse followup and Census 2000 unclassified households. In Census 2000, addresses that did not respond to the mailout had to be enumerated by nonresponse followup procedures. NRFU addresses are the most expensive to enumerate and may represent the most vulnerable segment of Americans. The household-level analysis provided a preliminary look at the conditions under which households formed from administrative records could be used for conventional NRFU households, obviating the need for fieldwork in those cases.

Addresses that had the status "unclassified" in Census 2000 were those for which so little information was available that occupancy status had to be imputed, and, conditional on being imputed "occupied," the entire household, including characteristics, had to be imputed as well. This treatment of unclassified households was the subject of a lawsuit reaching the U.S. Supreme Court (*Utah v. Evans*), in which the plaintiffs objected to the imputation substituting for enumeration. Although the census methodology prevailed, the possibility of enumerating these types of addresses by administrative records might provide a useful alternative to traditional imputation. This section provides some information comparing administrative records enumeration and Census 2000 imputations for the Census 2000 unclassified households in the AREX test sites.

Special terminology

For this section, the term "census household" refers to the persons enumerated at an address in Census 2000. The term "AREX household" refers to persons at an administrative records address. "Household size" refers to the number of people in the housing unit. For convenience, these definitions are applied to vacant housing units, so that when a Census or AREX address contains no people, the housing unit is assigned a household size of zero. We use the term "imputed household" for unclassified addresses whose occupancy status and household characteristics have been imputed.

Pairs of addresses (AREX and Census) that were matched by computer or clerical processes are referred to as "linked" housing units. The term "linked households" is used when comparing the properties of people within linked housing units. The term "demographic match" is used when two households have the same age, race, sex, and Hispanic origin distribution.

Finally, the term "AREX data" is used for administrative data obtained from the Bottom-up operations (i.e., including DMAF linkage, clerical review and FAV). The term "Census data" is used for data obtained from the Census 2000 HDF file.

Descriptive analysis

The household-level evaluation used both descriptive analyses and multiple regression analysis to assess the coverage and accuracy of AREX households. Descriptive analyses were performed for linked households in all five AREX counties and for the Census 2000 NRFU and imputed households in the test sites. These analyses provided the following evaluations:

- Coverage by AREX of its intended universe by determining the number and proportion of Census 2000 addresses that were matched by AREX addresses;

- Characteristics of Census 2000 households associated with AREX/Census matched addresses;

- Comparison of AREX and Census 2000 distributions of household size and household demographic characteristics;

- Characteristics of AREX households associated with AREX/Census 2000 household-to-household comparisons, including such properties as the presence of a person in the household of a particular race or ethnicity, and the presence of a person with a characteristic that was imputed in AREX.

Prediction model

To learn more about the characteristics of administrative records households that match the census and to take a first look at the potential uses of administrative records data to substitute for some part of the nonresponse followup or unclassified households in a conventional census, a logistic regression model was developed with the AREX/Census 2000 linked households as the units of analysis. The functional form of the model is Logit (Match=1|x) = xβ where Match is a dichotomous dependent variable, x is a vector of regressors, and β is a vector of constants to be estimated. For each linked address, the dependent variable was defined as follows:

$$\text{Match} = \begin{cases} 1 & \text{if the fully crossed age} \times \text{race} \times \text{sex} \times \text{Hispanic origin distributions in the} \\ & \text{linked Census household match the AREX household;} \\ 0 & \text{otherwise.} \end{cases}$$

This measure was based on the distribution of personal characteristics within an address and not on matches of individual persons. An address in AREX and in the census that had exactly the same distributional characteristics but were composed of entirely different persons would still receive a match score of 1. The simpler dependent variable--1 if all persons were the same, 0 otherwise--was not used because the AREX operations did not provide for matching individuals from AREX and census enumerations. Considering that the age distribution is in 5-year groups, the match definition used would appear to provide a result very close to an exact person match.

The regressors include characteristics of AREX households and characteristics of the linked addresses, representing the kind of information that would be available were data from administrative records to be used in support of a conventional census.

Limitations

The principal limitations on the ability to link addresses and demographically match households stemmed from the same deficiencies of the AREX administrative records files discussed in previous sections. First, the administrative data extracts were taken a year or more before census day. This means that movers, births, deaths, immigration and emigration, new housing, abandoned and demolished housing were unaccounted for a period of 12 or more months prior to census day. Second, many children are unaccounted for in administrative records at the national level; and therefore, AREX 2000 had difficulty enumerating children, generally, and, specifically, by virtue of the time lag problem and the limited demographics available for children on the Numident file (Miller, Judson, and Sater, 2000). Third, the race measurement and reporting deficiencies of the administrative records and differences in race measurement between AREX and the census presented serious challenges to comparisons matching race and Hispanic origin between members of AREX and census households. Finally, virtually all persons identified as having Hispanic origin in the AREX were imputed as such thus weakened the comparisons.

The AREX FAV had little impact on the household-level analysis; and a 100 percent FAV, if actually carried out, would have had little impact as well. Persons at administrative records addresses that would have been completely lost to the AREX as a result of the FAV would have had no impact on the household-level analysis since none of their addresses match the DMAF. And, as discussed in Section 2, there were would have been very few persons who remained in the enumeration but at different addresses as a result of the FAV.

Finally, deficiencies in administrative records and HDF addresses (for example, address duplication) and address matching technology resulted in a number of cases in which more than one administrative record address matched the same HDF address and vice versa. All of the administrative records addresses that matched the HDF but not on a one-to-one basis were excluded from the analyses.

Descriptive Analysis

AREX and Census address linkage

In the five counties covered by the experiment, the Census 2000 HDF contained 1,092,460 housing units (HUs) and 1744 group quarters (GQs), the latter excluded from this analysis. 24,584 (2.3 percent) of census households were "imputed households," and 360,914 (33.0 percent) were in the Census 2000 NRFU universe.

Of the 1,065,031 AREX addresses 992,865 were linked with addresses that existed in Census HDF; but 103,227 of the AREX addresses did not have a one-to-one link and were also excluded. This left 889,638 linked AREX addresses available for the household-level analysis. They represented 81.4 percent of census addresses.

Table 18 provides data on overall address linkage. AREX housing units (i.e. addresses) were linked with 84.0 percent of the 1,017,273 occupied census housing units. AREX housing units were linked with 46.4 percent of the 75,187 vacant census housing units. About 88 percent of AREX vacant units were found to be occupied by the census[10]. This confirms the discussion in Section 2 in which it was suggested that the AREX vacant addresses should have been canvassed as part of the Bottom-up process.

[10] Recall that AREX vacant housing units are those with an address that was linked to the HDF but for which no persons remained after best address selection.

Table 18. Coverage by AREX of Census housing units.

	Total	Linked with AREX housing units (% of total)	Linked with AREX occupied housing units (% of total)	Linked with AREX vacant housing units (% of total)
Census housing units	1,092,460	889,638 (81.4%)	813,688 (74.5%)	75,950 (7.0%)
Occupied Census housing units	1,017,273	854,741 (84.0%)	787,802 (77.4%)	66,939 (6.6%)
Vacant Census housing units	75,187	34,897 (46.4%)	25,886 (34.4%)	9,011 (12.0%)

AREX's coverage of the Census NRFU universe was not as good as its coverage of the non-NRFU universe. AREX housing units were linked with 70.9 percent of the 360,914 Census NRFU housing units, compared with 88.4 percent of the Census non-NRFU housing units. For occupied NRFU housing units, the coverage rate goes up to 76.7 percent. Table 19 contains more details about AREX's coverage of Census NRFU and non-NRFU housing units.

Table 19. Coverage by AREX of Census housing units, by NRFU status.

Type of Census housing unit	Total	Linked with AREX housing units	Linked with AREX occupied housing units	Linked with AREX vacant housing units
NRFU	360914	70.9%	60.8%	10.1%
non-NRFU	716450	88.4%	82.9%	5.5%
Occupied NRFU	289224	76.7%	67.1%	9.6%
Occupied non-NRFU	715115	88.5%	83.0%	5.5%
Vacant NRFU	71690	47.6%	35.2%	12.3%
Vacant non-NRFU	1335	58.7%	46.3%	12.4%

* Excludes 15,096 housing units in Census HDF with unknown NRFU status.

There were 24,584 imputed census housing units in the AREX test sites. AREX housing units were linked with 62.3 percent of them. AREX addresses were linked with 63.2 percent of those that were imputed to have people in them, and 34.7 percent of those imputed to be vacant. The linkage of imputed occupied units was about twice that of imputed vacant units, providing face validity for the Census 2000 imputation.

Table 20. Coverage by AREX of Census housing units, by imputation status.

Type of Census housing unit	Total	Linked with AREX housing units	Linked with occupied AREX housing units	Linked with vacant AREX housing units
Imputed	24,584	62.3%	51.7%	10.5%
Non-imputed	1,067,876	81.9%	75.0%	6.9%
Imputed occupied	23,811	63.2%	52.6%	10.6%
Non-imputed, occupied	993,462	84.5%	78.0%	6.5%
Imputed vacant	773	34.7%	25.5%	9.2%
Non-imputed, vacant	74,414	46.5%	34.5%	12.0%

The coverage by AREX of NRFU housing units and imputed housing units is not as good as for non-NRFU and non-imputed housing units. This may be due to several factors: (1) components of addresses from NRFU and/or imputed housing units might be generally of lower quality, and thus harder to match; (2) addresses of these housing units may be of types that are harder to match, e.g., those in apartment buildings, those on Rural Routes, or at P.O. boxes; and (3) people in these housing units may be more likely not to show up on any of the administrative records used for AREX.

AREX and Census household size

All occupied households

Table 21 shows the distributions of household size for linked and non-linked occupied households in AREX and for Census. The AREX distribution of household size was quite similar to the census distribution.

Table 21. Distributions of household size for Census and AREX for all five AREX counties. Occupied housing units only.

	Census		**AREX**	
Household	**Total**	**%[1]**	**Total**	**%[2]**
1	276590	27.2%	246726	27.9%
2	331472	32.6%	262075	29.6%
3	171136	16.8%	155929	17.6%
4	142822	14.0%	127295	14.4%
5	60988	6.0%	56596	6.4%
6	21655	2.1%	22695	2.6%
7-9	11275	1.1%	12481	1.4%
10+	1335	0.1%	1625	0.2%
All Sizes	**1,017,273**	**100%**	**885,422**	**100%**

[1] Percent of all Census occupied housing units

[2] Percent of all AREX occupied housing units

One salient feature of the data was that among the unlinked housing units in both Census and AREX, a very high percentage had one person. One possible explanation of this fact is that a much higher percentage of one-person households were at basic street addresses at which there are multiple housing units, and addresses at such basic street addresses (BSAs) were harder to link.

Linked occupied and non-occupied households

AREX and Census 2000 counted the same number of people in the housing unit for 51.1 percent of the 889,638 linked households, and AREX was within one of the census for 79.4 percent of the units. The 51.1 percent is effectively a ceiling on the percent of linked households that had exactly the same persons from AREX and Census 2000. Although errors in address linkage would account for some of the mismatched households, the deficiencies in administrative records cited earlier in this report--missing children, lack of special population operations and the time gap between the administrative records extracts and census day--most likely account for the major part.

For linked NRFU housing units, AREX had the same numbers of persons for 37.0 percent of the units and was within one 69.3 percent of the time. Evidently, Census 2000 NRFU housing units are more susceptible to AREX deficiencies than non-NRFU units. In addition, enumeration errors (such as “curbstoning”) in Census 2000 may be higher for these units than for units that responded to the initial mailout.

For the 15,043 linked imputed occupied households, AREX had the same count for 31.8 percent, and was within one for 66.8 percent of these addresses. The low percentage of household-by-household agreement between AREX and the census for imputed households should be expected from the error introduced by the imputation.

Table 22. Comparison of Census and AREX household size, by NRFU status, and by imputation status. For linked housing units.

AREX person count compared with Census	All Census housing units	Census non-NRFU housing units	Census NRFU housing Units	Non-imputed Census housing units	Imputed vacant Census housing units	Imputed occupied Census housing units
Same count	454,437	359818	94619	449,582	71	4,784
	(51.1%)*	(56.8%)	(37.0%)	(51.4%)	(26.5%)	(31.8%)
AREX one higher than Census	124,706	84269	40437	122,519	95	2,092
	(14.0%)	(13.3%)	(15.8%)	(14.0%)	(35.5%)	(13.9%)
AREX one lower	127,531	85178	42353	124,355	0	3,176
	(14.3%)	(13.4%)	(16.5%)	(14.2%)		(21.1%)
AREX 2 or 3 higher	64,635	36769	27866	63,024	77	1,534
	(7.3%)	(5.8%)	(10.9%)	(7.2%)	(28.7%)	(10.2%)
AREX 2 or 3 lower	79,848	47938	31910	77,463	0	2,385
	(9.0%)	(7.6%)	(12.5%)	(8.9%)		(15.9%)
AREX 4 or more higher	15,781	6486	9295	15,316	25	440
	(1.8%)	(1.0%)	(3.6%)	(1.8%)	(9.3%)	(2.9%)
AREX 4 or more lower	22,700	13158	9542	22,068	0	632
	(2.6%)	(2.1%)	(3.7%)	(2.5%)		(4.2%)
TOTAL	889,638	633,616	256,022	874,327	268	15,043
	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)

* Percents are percents of column total

Demographic comparisons of occupied linked households of the same size

In this section, demographic characteristics of linked households are compared. Because comparisons within households of different sizes are difficult to interpret, only linked occupied housing units in which AREX and Census 2000 have the same number of people are considered. There are 454,437 of these housing units representing 42.6 percent of all census housing units, 42.7 percent of all AREX housing units, and 51.2 percent of all linked housing units.

Tables 23-25 contain data only for linked households for which AREX and the census had the same total count. The tables show the frequencies with which AREX and the census agree for each:

Sex category;

Race category: White, Black, American Indian/Alaskan Native, Asian/Pacific Islander;

Hispanic origin category, i.e. Hispanic/non-Hispanic;

Five-year age category: 0-4, 5-9, …, 80-84, 85 and up;

Of the age categories: 0-17, 18-64, and 65 and up.

As expected, the agreements for racial composition and Hispanic origin composition were good -- in general, well above 90 percent. Generally household members tend to be all of one race and Hispanic origin. Also as expected, agreement rates did decline with household size because the likelihood of missing or AREX imputed race and different Hispanic origin imputations increases with number of persons in the household.

Agreement between AREX and the census across 5-year age groups provides an estimate of the proportion of households with exactly the same persons because it is improbable that two different households would agree in age distributions in 5-year categories. About 81 percent of the 445,426 households had the same 5-year category distribution. This is about 41 percent of all linked households.

The agreement rate for linked households of the same size is substantially higher for the age group distribution with only three categories, 0-17, 18-64 and 65 and up due to the increased tolerance for reporting errors and the greater probability of chance agreement.

Table 23. Comparisons between AREX and Census for demographic groups, for linked households (HH) with the same number of people only.

HH Size	Total linked, of equal size	Equal for all sex groups [1]	Equal for all race groups	Equal for all Hisp. groups	Equal for all 5-year age groups	Equal for age groups 0-17, 18-64, 65+	Equal for all demographic groups[3]
All sizes	445,426	91.2%[2]	93.4%	94.8%	81.3%	93.1%	80.5%
1	139,292	92.2%	95.1%	97.5%	82.5%	96.1%	85.4%
2	158,259	93.8%	94.8%	95.9%	83.9%	94.0%	84.3%
3	60,641	87.1%	90.7%	92.3%	75.7%	88.4%	72.2%
4	60,181	89.3%	90.7%	90.7%	80.8%	91.7%	74.0%
5	20,723	86.8%	88.9%	89.3%	77.2%	89.0%	69.5%
6	5,359	80.4%	86.0%	86.0%	68.0%	81.8%	59.2%
7+	971	56.8%	80.8%	83.0%	28.7%	52.7%	28.7%

[1] I.e., the AREX and Census households have the same number of males and the same number of females

[2] Percents are percents of the Total column

[3] Both sex groups, all race groups, both Hispanic origin groups, and age groups 0-17, 18-64, 65+

Table 24 shows that there was less AREX to Census 2000 agreement for NRFU households than for other census households, overall and controlling for size. Based on the 5-year age group match for Census NRFU households, only about 19 percent of AREX households linked with Census 2000 NRFU households seemed to have exactly the same persons. As expected there is even less agreement in household characteristics between AREX and Census imputed households (Table 25).

Table 24. Comparison of AREX and Census demographic composition of households. For linked households with the same number of people only, by size.

HH Size		Total	Equal for all sex groups[1,2]	Equal for all race groups	Equal for all Hisp. groups	Equal for all 5-year age groups	Equal for age groups 0-17,18-64, 65+	Equal for all demo-graphic groups[3]
All	NRFU	85,774	81.0%	87.7%	92.3%	58.1%	84.9%	63.4%
	non-NRFU	359,652	93.7%	94.7%	95.3%	86.9%	95.0%	84.6%
1	NRFU	31,313	82.5%	89.3%	95.7%	57.5%	91.1%	68.9%
	non-NRFU	107,979	95.0%	96.8%	98.1%	89.7%	97.5%	90.2%
2	NRFU	24,499	83.7%	88.5%	92.7%	58.6%	83.6%	64.9%
	non-NRFU	133,760	95.7%	96.0%	96.5%	88.6%	95.9%	87.9%
3	NRFU	12,549	75.7%	85.6%	89.4%	54.3%	77.1%	54.8%
	non-NRFU	48,092	90.1%	92.1%	93.0%	81.4%	91.4%	76.8%
4	NRFU	11,423	79.8%	86.3%	88.4%	63.2%	83.3%	60.2%
	non-NRFU	48,758	91.5%	91.7%	91.2%	84.9%	93.7%	77.3%
5	NRFU	4,473	78.1%	84.9%	87.2%	60.4%	80.0%	56.8%
	non-NRFU	16,250	89.2%	90.1%	89.9%	81.8%	91.4%	73.0%
6	NRFU	1,269	71.0%	80.4%	83.0%	54.0%	73.1%	46.8%
	non-NRFU	4,090	83.4%	87.8%	86.9%	72.4%	84.6%	63.0%
7+	NRFU	248	53.6%	79.8%	81.2%	27.0%	47.6%	24.6%
	non-NRFU	723	58.0%	81.2%	80.0%	29.3%	54.5%	30.2%

[1] I.e., the AREX and Census households have the same number of males and the same number of females.

[2] Percents are percents of Total.

[3] Both sex groups, all race groups, both Hispanic origin groups, and age groups 0-17, 18-64, 65+.

Table 25. Comparison of AREX and Census demographic groups within households. For linked households with the same number of people only, by size.

HH Size		Total	Equal for all sex groups	Equal for all race groups	Equal for all Hisp. groups	Equal for all 5-year age groups	Equal for age groups 0-17,18-64, 65+	Equal for all demo-graphic groups
All	Imputed	4,784	49.6%	74.9%	91.7%	7.0%	60.7%	23.0%
	Not imputed	440,642	91.7%	93.6%	94.8%	82.1%	93.4%	81.2%

Factors associated with demographic match rates

Single- and multi-unit BSAs

Table 26 contains data regarding comparisons of coverage rates, household size, and demographic characteristics for single- and multi-unit BSAs.

For all census household sizes, AREX addresses were less likely to link with census multi-unit addresses than with single-unit addresses. For linked households of equal size, AREX differed from census in all demographic groups more often for households at multi-unit addresses. The difference in percentage of demographic agreement is about 12 percent for households of size 1 and in the neighborhood of 20 percent for households of sizes greater than 1. Deficiencies in administrative records coverage and timing of the extracts most likely explain the differences in demographic agreement.

Table 26. Comparison of match rates and household comparisons between occupied housing units at multi-unit BSAs and housing units at single-unit BSAs.

Census HH Size	Group	Total	Linked (% of Total)	Equal size (%)[1]	Equal in all demographic groups (%)[2]
All sizes[3]	In multi-unit	278,447	188,826 (67.8%)	88,517 (46.9%)	64,992 (73.4%)
	In single-unit	738,826	665,915 (90.1%)	356,909 (53.6%)	293,720 (82.3%)
1	In multi-unit	135,833	91,051 (67.0%)	57,218 (62.8%)	44,978 (78.6%)
	In single-unit	140,757	125,568 (89.2%)	82,074 (65.4%)	74,034 (90.2%)
2	In multi-unit	80,719	55,820 (69.2%)	21,788 (39.0%)	15,009 (69.3%)
	In single-unit	250,753	226,676 (90.4%)	136,471 (60.2%)	118,386 (86.7%)
3-4	In multi-unit	51,244	35,165 (68.6%)	8,567 (24.4%)	4,459 (52.0%)
	In single-unit	237,644	237,644 (90.5%)	112,255 (47.2%)	83,906 (74.7%)
5-6	In multi-unit	9390	6,063 (64.6%)	926 (15.3%)	456 (49.2%)
	In single-unit	73,253	65,838 (89.9%)	25,156 (38.2%)	17115 (68.0%)
7+	In multi-unit	1,261	727 (57.7%)	18 (2.5%)	0
	In single-unit	11,349	10,189 (89.8%)	953 (9.4%)	279 (29.3%)

[1] Percent of linked

[2] Percent of linked of equal size

[3] Except size zero

Age of household occupants

The discrepancies between AREX and the census were due partly because some households have moved out of, and others moved into, addresses between the time of the administrative records cutoffs and the census. It is possible that households containing only older people are less likely to move, and may yield better AREX to the census comparisons. Table 27 provides address linkage rates by whether the housing unit is at multi-unit BSA, and by whether it has only people 50 and over. Table 28 provides comparisons of linkage rates, size, and demographics for housing units containing only people 50 and over, and others. (Tables B.16 and B.17A-B in Judson and Bauder (2002) contain similar comparisons for ages 18 and over, and for 65 and over.)

The coverage by AREX of census households with everyone over 50 was slightly, but consistently, higher. This was true whether controlling for multi-units or controlling for size. The comparison for household size and demographics were much better for one and two person households with all members 50 and over. The demographic comparison was worse for households of size 3 or more, but there were few of those where all members were 50 and over.

Table 27. Coverage by multi vs. single unit, and by household age characteristics.

Type of housing unit	Census household age characteristic	Total	Percent linked
All HUs	All 50 or older	292,091	85.8%
	Some under 50	639,088	79.9%
In multi-unit	All 50 or older	81,480	69.8%
	Some under 50	230,883	62.5%
In single-unit	All 50 or older	210,661	91.8%
	Some under 50	569,486	86.9%

Table 28. AREX to Census comparisons by size of housing unit and by household age characteristics.

Size of HH	Census household age characteristic	Total	Linked with AREX housing units (% of Total)	Equal size (%)[1]	Equal in all demographic groups[2] (%)[3]
1	All 50 or over	148,335	121,781 (82.1%)	86,518 (71.04%)	78,500 (90.7%)
	Some under 50	128,235	94,838 (74.0%)	52,774 (55.7%)	40,512 (76.8%)
2	All 50 or over	137,758	123,412 (89.6%)	83,662 (67.8%)	76,685 (91.7%)
	Some under 50	193,714	159,084 (82.1%)	74,597 (46.9%)	56,800 (76.1%)
3+	All 50 or over	5878	5,357 (91.1%)	2542 (47.5%)	2072 (81.5%)
	Some under 50	403,233	350,269 (86.9%)	145,313 (41.5%)	136,147 (93.7%)

[1] Percent of linked households

[2] Equal in: both sex groups, all race groups, both Hispanic origin categories, and age groups 0-17, 18-64, 65+

[3] Percent of linked of equal size

Race and Hispanic origin of household occupants

Table 29 shows how coverage, size comparisons, and race comparisons, vary with whether there was a person with race other than White in the household according to Census 2000.

For census households with at least one person other than White, the coverage by AREX is smaller, but not smaller by much, compared with households all of whose members were White. On the other hand, the household size comparisons and the racial composition comparisons display more disagreement for households with at least one person other than White. To some extent, this may be a consequence of race imputation that would have affected comparison of households with one or more persons other than White more often than all White households.

Table 29. The effect of the presence of Persons other than White in a household on household match rates and comparisons.

Census HH Size	Household type	Total	Linked with AREX housing units (% of Total)	Equal size (%)[1]	Equal in all four race groups (%)[2]
All sizes	All White	740,218	631,606 (85.3%)	358,833 (56.8%)	347,592 (96.9%)
	At least one Other race	278,799	223,135 (80.0%)	86,593 (38.8%)	68,356 (78.9%)
1	All White	205,226	165,098 (80.5%)	111,112 (67.3%)	108,450 (97.6%)
	At least one Other race	71,498	51,121 (72.1%)	28,180 (54.7%)	24,049 (85.3%)
2	All White	256,585	221,806 (86.5%)	133,180 (60.0%)	130,033 (97.6%)
	At least one Other race	75,038	60,690 (80.9%)	25,.079 (41.3%)	19,995 (79.7%)
3-4	All White	219,030	192,772 (88.0%)	93,694 (48.6%)	89,491 (95.5%)
	At least one Other race	95,207	80,037 (84.1%)	27,128 (33.9%)	20,105 (74.1%)
5-6	All White	53,202	46,707 (87.8%)	20,319 (43.5%)	19,144 (94.2%)
	At least one Other race	29,635	25,194 (85.0%)	5,763 (22.9%)	3,896 (67.6%)
7+	All White	6,175	5,223 (84.6%)	528 (10.1%)	474 (89.8%)
	At least one Other race	7,421	5,693 (76.7%)	443 (7.8%)	311 (70.2%)

[1] Percent of linked households

[2] Percent of linked households of equal size

AREX coverage of census addresses did not differ much between households with and without Hispanics (Table 30). However, households with one or more Hispanics in the census were much less likely to match corresponding AREX households in size and Hispanic/non-Hispanic composition. Differences in household sizes were most likely due to deficiencies in administrative records coverage of Hispanics and the age of the records vis-à-vis the Census. Differences in Hispanic composition within households of equal size were most likely due to the fact that Hispanic origin was model-based for virtually all AREX persons.

Table 30. The effect of the presence of Hispanics on household match rates.

Census HH Size	Household type	Total	Linked with AREX housing units (% of Total)	Equal size (%)[1]	Equal # of Hispanics (%)[2]
All sizes	All Nonhispanic	956,474	803,272 (84.0%)	424,867 (52.9%)	411,698 (96.9%)
	At least one Hispanic	62,533	51,469 (82.3%)	20,559 (39.9%)	10,365 (50.4%)
1	All Nonhispanic	269,018	210,745 (78.3%)	136,114 (64.6%)	134,063 (98.5%)
	At least one Hispanic	7,706	5,874 (76.2%)	3,178 (54.1%)	1,802 (56.7%)
2	All Nonhispanic	314,587	268,371 (85.3%)	151,588 (56.5%)	147,697 (97.4%)
	At least one Hispanic	17,036	14,125 (82.9%)	6,671 (47.2%)	4,053 (60.8%)
3-4	All Nonhispanic	287,966	250,589 (87.0%)	112,467 (44.9%)	106,922 (95.1%)
	At least one Hispanic	26,271	22,220 (84.6%)	8,355 (37.6%)	3,609 (43.2%)
5-6	All Nonhispanic	73,654	64,212 (87.2%)	23,831 (37.1%)	22,235 (93.3%)
	At least one Hispanic	9,183	7,689 (83.7%)	2,251 (29.3%)	876 (38.9%)
7+	All Nonhispanic	11,249	9,355 (83.2%)	867 (9.3%)	781 (90.1%)
	At least one Hispanic	2,347	1,561 (66.5%)	104 (6.7%)	25 (24.0%)

[1] Percent of linked

[2] Percent of linked of equal size

AREX race imputation

Table 31 concerns linked households in which no person's AREX race was imputed, and those in which at least one person's race was imputed. The comparison was done with regard to the racial composition of the household. As expected, households with imputed race were less likely to agree on household race composition. Although the overall agreement rate of 86 percent was quite high when one or more members had imputed race, the agreement rate may have been much smaller when a member was imputed to be of an other race.

Table 31. The effect of AREX imputed race on household comparisons.

		HHs with at least one person with AREX imputed race		HHs with no person with AREX imputed race	
Census HH Size	Total linked, with equal size	Number (% of [1])	Equal in all race categories	Number (% of [1])	Equal in all race categories
	[1]	[2]	(% of [2])	[3]	(% of [3])
All sizes*	445,426	100,416 (22.5%)	86,290 (85.9%)	345,010 (77.5%)	329,658 (95.6%)
1	139,292	5,197 (3.7%)	4,099 (78.9%)	134,095 (96.3%)	128,400 (95.8%)
2	158,259	14,087 (8.9%)	11,351 (80.6%)	144,172 (91.1%)	138,677 (96.2%)
3-4	120,822	61,389 (50.8%)	53,689 (87.5%)	59,433 (49.2%)	55,907 (94.1%)
5+	27,053	19,743 (73.0%)	17,151 (86.9%)	7,310 (27.0%)	6,674 (91.3%)
5-6	26,082	18,991 (72.8%)	16,558 (87.2%)	7,091 (27.2%)	6,482 (91.4%)
7+	971	752 (77.4%)	593 (78.9%)	291 (22.6%)	192 (87.7%)

* Not including zero

Predicting AREX/Census household similarity

The purpose of the regression analysis was to try to understand more about those circumstances under which AREX administrative records households would match census households in both number and demographic composition. This would also provide a first look at the potential uses of administrative records data to substitute for some part of the nonresponse followup or unclassified households in a conventional census.

Model Specification

For this initial model-building attempt, the units of analysis were all 889,638 one-to-one linked households. Separate equations for Census 2000 NRFU and unclassified households were not estimated, but dummy variables were included in the equation for these two types of decennial census household outcomes to see whether other predictor variables had accounted for differences in AREX/Census 2000 household similarity noticed in the descriptive analyses. This approach assumed that the regression hyperplanes for the three types of census households were parallel, an assumption that will be tested in future analyses.

Also, the analysis reported here attempted to account for household size agreement and demographic composition simultaneously, providing in some sense, the net association between predictors and outcomes. However, it is possible that associations between predictors and household size agreement are different from associations with demographic similarity given agreement in size. This possibility will also be explored in future work.

The functional form of the model is

$$y = \ln \frac{P(\text{Match} = 1 \mid x)}{P(\text{Match} = 0 \mid x)} = \text{Logit}(P(\text{Match} = 1 \mid x) = x\beta$$

where Match is a dichotomous dependent variable, x is a vector of regressors, and β is a vector of constants to be estimated. For each linked address, the dependent variable was defined as follows:

$$\text{Match} = \begin{cases} 1 & \text{if the fully crossed age} \times \text{race} \times \text{sex} \times \text{Hispanic origin distributions in the} \\ & \text{linked Census household match the AREX household;} \\ 0 & \text{otherwise.} \end{cases}$$

Predictor variables

The regressors, x, include characteristics of AREX addresses and households that would be available were data from administrative records to be used in support of a conventional census. The variable is dichotomous, taking on the value 1 if the characteristic is present and 0, otherwise. The interaction terms are products of the individual predictors. The predictors are numbered with the "Row num" in Table 34. The predictors were chosen after examination of an extensive set of bivariate crosstabulations with the dependent variable. The tabulations and associated discussion can be found in Judson and Bauder (2002).

Address administrative records source files

Generally, it was assumed that addresses appearing in more than one administrative record source file would be less likely to represent a moving household than addresses found in only one file. Additionally households with addresses in Medicare files would largely represent older persons and represent stable households. The following variables pertain to the source of the "best" administrative records address.

[10] **In IRS file** -- In the IRS 1040 file.

[11] **In IRMF file** -- In the IRS Information Returns Master File (i.e. the 1099 file).

[12] **In Medicare file** -- In the Medicare eligibility file

[13] **In IRS & IRMF** --[10]*[11]

[14] **In IRS & MED** -- [10]*[12]

[15] **In MED and IRMF** -- [11]*[12]

AREX Household characteristics

The following are characteristics of AREX households thought to be associated with same size and demographic similarity with the linked census households. To a certain extent, these variable are suggested by the descriptive analysis of the previous section, keeping in mind that the descriptive analysis often used census household characteristics rather than AREX household characteristics.

[5] **One or two persons** -- Household contains only 1 or 2 persons.

[6] **No imputed race** -- No household member has imputed race.

[7] **Hhold has children** -- Household has one or more children under the age of 18.

[8] **Hhold has 1+ White** -- Household has one or more White members.

[9] **Hhold all age 65+** -- All members of the household are age 65 or over.

[16] **Age 65+ & One/two** -- [5]*[9]

[17] **Age 65+ & 1+ White** -- [8]*[9]

[18] **One/two & 1+ White** -- [5]*[8]

[19] **65+ & 1 or 2 & 1+ White** -- [5]*[8]*[9]

[21] **65+ and no imputed race** -- [6]*[9]

AREX/ DMAF address characteristics

Single unit addresses are assumed to be more predictable than multi-unit addresses.

[4] **Not multi unit** – AREX indicates that the address is single unit.

[20] **65+ and Not multiunit** -- [4]*[9]

[22**] No imputed Race and not multi** -- [4]*[6]

[23] **65+ & No imp. & not multi** -- [4]*[6]*[9]

[1] **Colorado effect** -- AREX address is in Colorado

Census 2000 response type

Including variables representing Census 2000 response type provides a first indication of whether separate models might be needed for each type. (The reference group is mailback respondents.)

[2] **Enumerator return** -- NRFU respondent household

[3] **Imputed return** -- Census 2000 whole house imputation

Regression results

Although the regression results are useful in obtaining an initial understanding the relationships between AREX address and household characteristics and AREX/Census Match status, it is important to keep in mind that these matched households are not a nationally representative sample, that the analysis is exploratory in nature, and that improvements in administrative records processing in the future will be substantial. Therefore, these results should be considered illustrative in nature.

Table 32. Overall Response Profile for the "Match" Variable

Response Profile and Overall Model Fit Statistics	
Match Status	Total Frequency
Demographics Match	342294 (38.5%)
Demographics Do Not Match	547344 (61.5%)

About 38.5 percent of all linked addresses also matched on demographics.

Table 33. Goodness-of-Fit Measures for the Logistic Regression Model

Criterion	Intercept Only	Intercept and Covariates
AIC	1,185,613.2	1,001,550.2
SC	1,185,624.9	1,001,831.0
-2 Log L	1,185,611.2	1,001,502.2
Pseudo R-Square	0.1869	

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	184,108.945	23	<.0001

(full model versus null model of intercept only)

Note: N=889,638 households in two AREX test sites in Colorado and Maryland whose addresses were computer linked; A household is declared "matched" if it's age, race, sex and Hispanic origin composition is the same across the AREX household and the equivalent census household. AIC is the Akaike Information Criterion; SC is the Schwarz criterion. –2 Log L is –2 times the log likelihood (LL) of the model, evaluated at its maximum; R-square is the pseudo R-square value, consisting of (LL(model) – LL(intercept only))/LL(model). The Likelihood Ratio test tests the null hypothesis that all coefficients except the intercept are zero in the population; Pr>ChiSq is the (nominal) probability of obtaining that Chi-Square value by chance; Because observations may not be I.I.D., standard errors may be understated and significance levels overstated. (Note is also applicable to Table 34).

All of the variables taken together significantly improved the prediction of Match status. The Pseudo R-Square value indicates that the model results in a 19 percent improvement in the log-likelihood over the null model of an intercept only.

Coefficient estimates

Table 34 provides maximum likelihood estimates of the coefficients for the full model. The rightmost column indicates exponentiated coefficients, and can be interpreted as the (multiplicative) change in the odds of being a match given the corresponding characteristic, holding all other variables constant. An exponentiated coefficient of "1" indicates no effect, greater than 1 indicates positive effect, and less than 1 indicates negative effect.

The presence of interaction terms makes the interpretation of individual coefficients somewhat difficult. Still, the results for AREX address and household characteristics seem generally as expected. AREX households that are smaller (one or two members), have only members aged 65+, have one or more Whites, and have no members with imputed race tend to be more likely to match the corresponding Census 2000 household. AREX households at single-unit addresses are more likely to match the census than those at multi-unit addresses.

The negative coefficients on the Enumerator Return and Imputed Return indicate that these households remain less predictable other factors held constant. Separate equations for Census NRFU households might be required. Households where the census return was imputed are very unlikely to have the same demographics as their AREX counterparts and have added some noise to the coefficient estimates.

Table 34. Maximum Likelihood Parameter Estimates, Standard Errors, and approximate tests

Row Number	Variable	df	Estimate	Standard Error	Wald Chi-Square	PR >ChiSq	Exp (Est)
[0]	Intercept	1	-2.756	0.050	2977.43	<.0001	0.064
[1]	Colorado Effect	1	-0.102	0.005	379.62	<.0001	0.903
[2]	Enumerator Return	1	-1.096	0.006	26648.72	<.0001	0.334
[3]	Imputed Return	1	-3.133	0.110	809.52	<.0001	0.044
[4]	Not Multi-unit	1	0.926	0.018	2656.05	<.0001	2.525
[5]	One or Two Persons	1	0.982	0.011	7013.33	<.0001	2.672
[6]	No Imputed Race	1	0.790	0.018	1778.60	<.0001	2.205
[7]	Hhold has Children	1	0.275	0.007	1239.27	<.0001	1.317
[8]	Hhold has 1+White	1	0.598	0.009	4168.03	<.0001	1.819
[9]	Hhold all age 65+	1	0.281	0.187	2.25	0.1334	1.325
[10]	In IRS File	1	-0.048	0.047	1.04	<0.3075	0.953
[11]	In IRMF File	1	-0.341	0.047	52.61	<.0001	0.710
[12]	In Medicare File	1	-0.076	0.048	2.50	<0.1136	0.927
[13]	In IRS & IRMF	1	0.901	0.047	363.32	<.0001	2.462
[14]	In IRS & Medicare	1	-0.488	0.015	996.77	<.0001	0.614
[15]	In Medicare and IRMF	1	0.390	0.047	68.23	<.0001	1.478
[16]	Age 65+ & One/Two	1	0.870	0.156	30.81	0.0001	2.389
[17]	Age 65+ & 1 + White	1	-1.042	0.167	38.63	<.0001	0.353
[18]	One/Two & 1 + White	1	-0.036	0.013	8.001	<0.0047	1.037
[19]	65+ & 1 or 2 & 1+ White	1	0.974	0.168	33.25	<.0001	2.649
[20]	65+ and not Multi-unit	1	-1.021	0.119	73.41	<.0001	0.360
[21]	65+ and no Imputed Race	1	0.425	0.105	16.23	<.0001	1.531
[22]	No imputed race and not Multi	1	-0.630	0.019	1057.22	<.0001	0.532
[23]	65+ & no imputed Race & not Multi	1	0.657	0.120	29.90	<.0001	1.931
[10]*[11]*[13]	Total Effect of Capture in IRS and IRMF	(given not in Medicare)					1.666
[10]*… …*[15]	Total Effect of Capture in all Three Files	(w/o three-way interaction)					1.401
[5]*[8]* [9]*[16]… *[19]	Total Effect of all of 65+, White, and 1/2 Person Hhold						14.92
[4]*[6]* [9]*[20] …*[23]	Total Effect of all of 65+, Nonmulti-unit, nonimputed race						4.177

The last four rows of the table indicate the net effect of some combinations of variables, calculated by multiplying their exponentiated coefficients. For example, the total effect of being captured in IRS, IRMF, and Medicare however, is effectively that the household is about 1.6 times more likely to demographically match.

The effect of having all persons 65 or older, at least one White person, and consisting only of a one or two person household (given that the household is multi-unit and has at least one member with imputed race) is dramatically positive, 14.92. Similarly, a household having all persons 65 or older, not being a multiunit address, and having no imputation from the administrative records (but also other than white and more than two persons) is about four times more likely to match census demographics, holding other effects constant.

Goodness of Fit

One way to evaluate the ability of the model to correctly predict household match status is to establish a decision rule that first chooses a probability level, c, and then deems an observation to be demographically matched if the probability that Match = 1 for that observation, calculated from the model, is greater than c. More succinctly, for a given level of c, $0 \leq c \leq 1$, if $P(Match=1|x\beta) \geq c$, predict "AREX household is demographically matched." Otherwise predict that the household is not demographically matched. For a given probability level, there are several measures that can be used to evaluate the decision rule.

Accuracy: Proportion of all cases correctly classified.

False positive: Proportion of cases where the true match status is 0 given that the prediction is 1.

False negative: Proportion of cases where the true match status is 1 given that the prediction is 0.

Sensitivity: Proportion of cases where the prediction is 1 given that the true match status is 1.

Specificity: Proportion of cases where the prediction is 0 given that the true match status is 0.

Table 35 shows the estimates of these quantities for decision rule probability levels between .5 and .9.

Table 35. Classification Results for Predicted Probabilities.

Classification Table									
	Correct		Incorrect		Percentages				
Prob. Level	Event	Non-Event	Non-Event	Event	Correct	Sensit-ivity	Specif-icity	False POS	False NEG
0.5	184,230	457,943	89,401	158,064	72.2	53.8	83.7	32.7	25.7
0.6	110,701	506,699	40,645	231,593	69.4	32.3	92.6	26.9	31.4
0.7	72,335	530,307	17,037	269,959	67.7	21.1	96.9	19.1	33.7
0.8	32,373	540,798	6,546	309,921	64.4	9.5	98.8	16.8	36.4

As can be seen, if we choose the cutoff of 0.5 (so that we predict a "match" when P[Match=1|xβ] is greater than or equal to .5), we obtain about 184,000 correct match predictions, and about 458,000 correct nonmatch predictions. Dividing the sum of the correct predictions by the total number of cases above the 0.5 cutoff (about 889,000) gives 72.2 percent correct predictions (accuracy). Similarly, 54 percent of the matches were correctly predicted to be matches (sensitivity); 83.7 percent of the nonmatches were correctly predicted to be nonmatches (specificity); there was a 32.7 percent false positive rate and a 25.7 percent false negative rate.

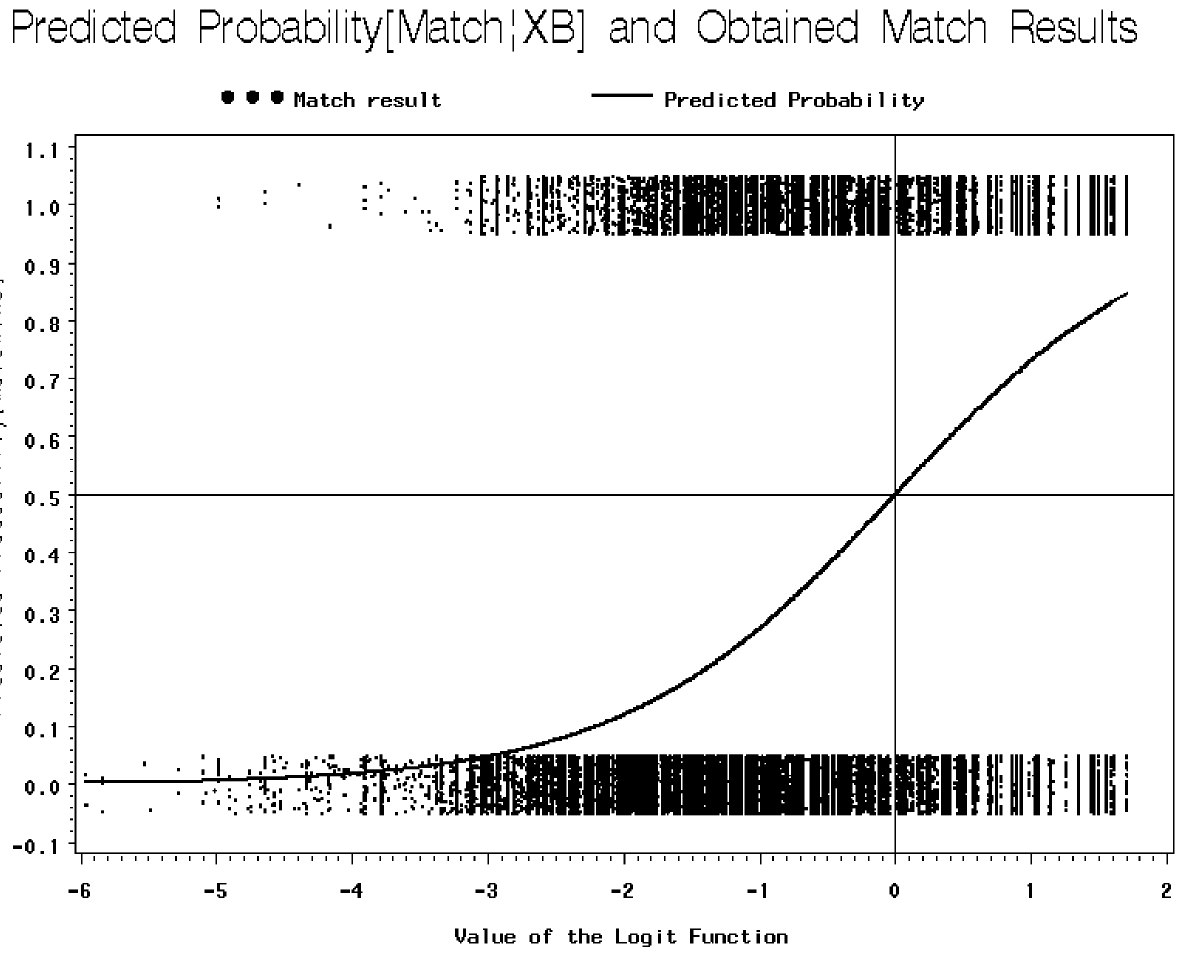
As the cutoff level, c, increases (i.e., becomes more stringent), the probability of making an error in deeming a match status of 1 above the cutoff (probability of a false positive) declines. For example at c=0.8, the number of correct predictions is 32,373 and the number of incorrect predictions is 6,546 for a total of 38,919. Thus the probability of a false positive is 6,546/38,919 = 0.168. As shown in the table, as c increases, the sensitivity and overall correctness decline, and specificity and the probability of a false negative increase.

In order to evaluate cutoffs and their implications for goodness of fit, sensitivity and specificity, we present the following evaluative figures. **Error! Reference source not found.** provides an assessment of the goodness of fit of the obtained logit function against "jittered" outcomes.

In this figure, the ordinate is the value of the logit function ln(p/1-p). A 10 percent sample of the 889,638 observations are plotted here. Each individual observation (a linked pair of addresses) is plotted as a point near zero or one. The points have been "jittered" slightly to simulate density and avoid overplotting. The abscissa is the predicted probability that an observation will be a match. If we choose 0.5 as our cutoff (so that we declare an observation a predicted match is P[match|XB]>0.5), then this corresponds to a logit value of zero, and the vertical line. The horizontal line at 0.5 is for reference. Points in the upper right hand quadrant are "hits"—correct predictions that the demographics of the households match. Points in the lower left hand quadrant are also "hits"—correct predictions that the demographics of the households will not match.

Points in the upper left hand and lower right hand quadrants are misses—incorrect predictions. Comparing the predicted logit function to the density of the obtained match outcomes assesses goodness of fit. (For more on the development and interpretation of this graph, see Judson, 1992. For more goodness of fit measures and graphics, see Judson and Bauder, 2002.)

Figure 18. Goodness of Fit Diagnostic Plot.



In thinking about using a regression model approach in deciding when to substitute an administrative records household for a nonresponse household in a conventional census, the probability of false match would have to be small, providing confidence that the household substitution was accurate and obviating the need for further enumeration. But the proportion of households in scope for substitution, that is, the proportion of households above the decision cutoff level would have to be large enough to provide substantial savings over face-to-face enumerations. For example, from Table 35, a cutoff of 0.8 would provide a relatively low probability of false negative, 0.168; but the proportion of households in scope for substitution at that cutoff would be about 4 percent (38,919/889,638).

Summary and conclusions

General similarity between AREX data and Census data

A summary of the AREX to Census 2000 comparisons is given in Table 36.

The overall coverage of occupied census housing units by AREX was about 84 percent (81 percent of occupied and vacant units). The coverage of census addresses by administrative records addresses could be raised substantially by resolving matches that were not one-to-one, by filling coverage gaps in administrative records, and by obtaining administrative records extracts at points in time closer to census day. Proposals for accomplishing all of these tasks are provided in Section 5 of this report.

Similarity in size and demographic composition of linked AREX and census households was rather low. Of the occupied census linked households, AREX and the census had the same number of people in 52.1 percent of the cases (51.4 percent of all linked households). In 41.9 percent of occupied linked households, AREX and the census had the same number of people, and the same demographic distributions using the three broad age categories.

About 81 percent of households had the same 5-year age distribution and about 93 percent had the same age distribution in the three broad groups. This suggests that the proportion of households of the same size that had exactly the same persons was somewhere in between. The relatively low percentage of households (80.5 percent) with similarity along all demographic dimensions was due in large part to race and Hispanic origin imputation and the difference in race categories between AREX and the census. It is unlikely that even improved race and Hispanic origin models will be sufficient, in themselves, for decennial census enumeration. Another approach is currently being developed. (See the discussion in Sections 4 and 5.)

In summary, the key deficiency of the AREX administrative records processing was the failure to get the right number of people (and, therefore, the right people) at many of the addresses. Dissimilarity of households is of special concern for an AREX-type of design because of the limited opportunities to correct that part of the enumeration obtained from the administrative records. This will be the biggest challenge for future administrative records development.

Table 36. Summary of match rates and household comparisons between AREX and Census.

Type of Housing Unit	All of Census	NRFU	non-NRFU	Imputed HHs	non-Imputed HHs
Total Occupied Census Housing Units	1,017,273	289,224	728,049	23,811	993,462
Census Occupied, linked	854,741 (84.0%)[1]	221,909 (76.7%)	632,832 (86.9%)	15,043 (63.2%)	839,698 (84.5%)
Linked occupied with equal number	455,426 (52.1%)[2]	85,774 (38.7%)	359,652 (56.8%)	4,784 (31.8%)	440,642 (52.5%)
AREX and Census counts both sex categories	406,349 (91.2%)[3]	69,488 (81.0%)	336,861 (93.7%)	2,373 (49.6%)	403,976 (91.7%)
AREX and Census counts equal in all race categories	415,948 (93.4%)[3]	75,262 (87.7%)	340,686 (94.%)	3,583 (74.9%)	412,365 (93.6%)
AREX and Census counts equal in both Hispanic origin categories	422,063 (94.8%)[3]	79,146 (92.3%)	342,917 (95.4%)	4,388 (91.7%)	417,675 (94.8%)
AREX and Census counts equal in all 5-year age categories	362,202 (81.3%)[3]	49,833 (58.1%)	312,369 (86.9%)	335 (7.0%)	361,867 (82.1%)
Equal in age groups 0-17, 18-64, 65+	414,668 (93.1%)[3]	72,835 (84.9%)	341,833 (95.1%)	2,905 (60.7%)	411,763 (93.5%)
AREX and Census counts equal in sex, race, Hispanic origin, and 5-year age groups	333,577 (74.9%)[3]	43,210 (50.4%)	290,367 (80.7%)	138 (2.9%)	333,439 (75.7%)
AREX and Census equal in demographic composition: sex, race, Hispanic origin, and age groups 0-17, 18-64, 65+	358,712 (80.5%)[3]	54,400 (63.4%)	304,312 (84.6%)	1,099 (23.0%)	357,613 (81.2%)

[1] Percent of Census occupied housing units

[2] Percent of Census linked housing units

[3] Percent of linked housing units with equal numbers of people

Similarity between AREX and Census NRFU and imputed households

There was less similarity between AREX and Census NRFU households than non-NRFU households across all outcome measures. The address linkage rate for Census NRFU households was about 77 percent compared to 84 percent for non-NRFU households. For NRFU households, AREX and census agreement on the household size was 39 percent (57 percent for non-NRFU), and agreement on all demographic groups given agreement on size was 63 percent (85 percent for non-NRFU).

These results suggest that substituting AREX households for NRFU households in a conventional census will be more difficult than matching households in general. It may be that households for which administrative records are weak overlap disproportionately with the Census NRFU population. However, it may also be the case that the characteristics of Census NRFU were more likely to be affected by AREX source file cutoffs and other AREX processing decisions than non-NRFU households. Also the Census 2000 enumeration of NRFU households may be less reliable.

Similarity between AREX administrative records households and Census 2000 unclassified households was substantially weaker than with the NRFU households. This is not surprising since the census was least sure of the status of these addresses and the persons placed at them were imputed by the Census. AREX 2000 did not provide the information needed to assess whether using administrative records to enumerate census unclassified households would be more accurate than the imputation.

Predicting household similarity

The logistic regression model predicted modestly well when AREX and census households matched demographically. Factors that predicted demographic matches included: one or two person households, households with exclusively older persons, households where members are captured by more than one administrative record system, households with no race imputation, and households that were at single-unit addresses.

A decision rule that deemed matches if predicted probability was above 50 percent resulted in correct match status in 72 percent of the cases but with a false positive rate of about 33 percent. The most stringent cutoff of 80 percent reduced the false positive rate to 16.8 percent, entailed only about 4 percent of the household population.

4. DISCUSSION AND RECOMMENDATIONS FOR 2010 PLANNING AND OTHER CENSUS BUREAU PROGRAMS

4.1 Additional AREX 2000 evaluations

Assess the net impact of clerical and field Processes on the Bottom-up enumeration

There were originally four operations in AREX 2000 that were designed to improve administrative records addresses. MAFGOR was used to geocode city-style addresses that were not coded by computer. The RFPA was designed to obtain physical addresses for persons whose mailing addresses were non-city style or P.O. Box and geocode them. The clerical reviews following the initial match to the MAF and the FAV were designed to validate administrative record addresses that did not match the MAF. All of these operations were complex and labor intensive.

It is not possible to evaluate each of the clerical and field operations separately because the AREX design did not vary these factors experimentally, and the RFPA results were not included in the AREX. Still it may be possible to gauge the net effect of the three operations on the Bottom-up results by stepping through the Bottom-up process excluding address information from any of the clerical or field operations and comparing the results to the Bottom-up enumeration.

The impact of eliminating the MAFGOR, clerical review of the first DMAF match, and the FAV would be threefold. First, the effect of eliminating the three operations would be to decrease the number of persons enumerated from administrative records. Those persons at addresses that did not computer geocode, did not match the MAF through computer operations, or were found to be valid only by the FAV would be eliminated if their addresses were all of these types. Second, selected addresses for some individuals enumerated from administrative records would change because their current Bottom-up address would be eliminated leaving some other address still acceptable to the Bottom-up process. Finally, because the total number of acceptable administrative record addresses would be smaller, there would be more non-matched DMAF records to canvass in order to complete the enumeration. That is, the number of addresses brought in by the Census Pull, simulating the canvassing, would be larger.

In broad terms, eliminating the results of the three address improvement operations would require the following steps:

- Recreate the address lists available to the Bottom-up by removing all addresses that were in the original list due to any of the three operations. (Because the new address list is a subset of the original list, an additional match to the DMAF for tabulation block codes should not be required. This ignores the possibility that a MAFGOR coded address might have picked up a block code from the DMAF);
- Recreate the Bottom-up composite person records by matching the smaller set of addresses to the individuals and reapplying the address selection rules; and,

- Rematch to the Census 2000 HDF in the AREX sites and include persons at unmatched HDF addresses.

There are a number of ways that this alternative process could be evaluated. First, the alternative Bottom-up counts could be compared with the Census 2000 using methodology similar to that of Section 3. The purpose would be to test if basic enumeration results were different from those of the original Bottom-up process.

Matching the two sets of Bottom-up results person by person could make a more extensive analysis. This would permit an analysis of (1) persons lost completely to the Bottom-up as a result of dropping the address operations, (2) persons whose address changed within the site, (3) persons omitted from administrative records counts who were picked up at additional Census Pull addresses, and (4) new Census Pull persons not in the first AREX Bottom-up results. Of course, a person-to-person match would require substantial work. The methodology used by Wagner (2002) in matching the Numident to the Census 2000 HCUF may provide a useful approach.

Finally, there could be analyses that focus on the addresses rather than the persons. Address analyses could address the administrative records sources of rejected addresses, the impact of rejected addresses on address selection, and a comparison of addresses for persons omitted from the administrative records counts but who showed up in the Census Pull. In the latter, it might be important to understand why the Census Pull address was not obtained from administrative records address selection.

Adding AREX vacants and unduplicating Bottom-up results

In addition to comparing the two Bottom-up processes described above, consideration should be given to creating two additional Bottom-up enumerations based on proposals offered in the Bottom-up evaluation in Section 3: (1) adding the vacant AREX addresses to the Census Pull if they matched the HDF, and (2) unduplicating individuals between the Census Pull and the administrative records. This pair of Bottom-up enumerations would appear to be more correct than the corresponding pair without these additional operations.

Repeating AREX 2000 with StARS 2000 without clerical or field operations

If eliminating the clerical address operations, as discussed in Section 2, turns out to be relatively inconsequential, then there would be great value in "repeating" the AREX Bottom-up process (without the clerical address operations) and the statistical evaluations with StARS 2000. The reason is that the administrative records data sets used in StARS 2000 are much closer to those that might be available in an actual administrative records census than those of StARS 1999 (putting aside the possibility of additional administrative records sources). Having the results from this administrative records baseline will be important in planning for AREX 2010 because the test items will not be confounded with the timing problems due to the administrative records extract dates in StARS 1999.

Since the purpose of redoing the Bottom-up would not be primarily to compare with the StARS 1999 results, operational improvements, such as improved SSN validation for IRS files that have been incorporated into StARS 2000 would be appropriate. The StARS 2000 AREX should include DMAF matched AREX vacant addresses in the Census Pull, and there should be unduplication after it. Also, the more timely administrative records would provide an opportunity to take a new look at the address selection algorithm, redirecting the emphasis to choosing the address that best reflects residence as of Census Day and away from the somewhat artificial focus on the presence of block codes.

Using the StARS 2000 files does not resolve two of the major limitations on the AREX: the handling of special populations, and race and Hispanic origin measurement. For the latter, consideration should be given to using Wagner's race and Hispanic origin data (2002). Although this is somewhat circular; again, it might provide results that are much closer to what would have been achieved in an actual census. Correct handling of special populations may need to await future experiments.

Analysis of administrative records coverage gaps[11]

In this report, a number of coverage gaps in administrative records for both adults and children have been identified; but the population sizes and characteristics of the missing persons is not known precisely. A linkage of StARS to a Survey of Income and Program Participation (SIPP) or Current Population Survey (CPS) sample for a corresponding time period would provide some information about the characteristics of persons in the survey who were not found in the administrative records. The linkage could be easily accomplished using the SSNs developed for those surveys, and using probabilistic methods for those without SSNs. The subsequent analysis would not provide a complete look at the non-covered population because of coverage deficiencies in the survey, itself, both in terms of segments of the population underrepresented and persons in the survey sample for whom SSNs are not available. Still, much could be learned about the adults and children not found in the administrative record systems.

A SIPP linkage might be particularly useful for missing adults because it would identify the government programs in which they are participating and provide details about social and economic circumstances. For households in which adults have been matched but not all of the children, a key focus might be to identify the administrative records characteristics of the adults who then might become an additional special population for administrative records census purposes. That is, these households might receive a special mailing in order to obtain a more complete enumeration of the household in AREX 2010.

[11] Suggestions for additional administrative records acquisitions are given in Bye, 2002, section 5.

4.2 Race and Hispanic Origin enhanced Numident

Improved modeling of race and Hispanic origin for administrative records will not provide a general solution to decennial census measurement for two reasons. First, it is unlikely that the models can provide adequate fit for the most difficult to measure groups such as American Indians and Alaskan Natives, Hawaii Islanders, and persons classifying themselves as multi-race. Second, even the best models will always suffer from a lack of fit in small geographic areas due to variation in local responses about model predicted averages.

Looking ahead to 2010, it is important to continue to annotate the Numident with survey reported race and Hispanic origin in order to make the annotated Numident as complete as possible. The American Community Survey (ACS) would be a major source of ongoing updates, under full implementation.

Because there will always be a residual subset of the Numident for which survey responses to race and Hispanic origin are not available, there will continue to be a need for race and Hispanic origin models. The methods proposed here make that subset smaller and smaller.

The initial task of annotating the Census Numident with race and Hispanic origin from Census 2000 was described briefly in Section 4, and an ongoing process to fill in the remaining gaps in the Numident was proposed. There are some research activities that should be considered in connection with this process.

Evaluation of the initial match

First, the accuracy of the Numident/Census match should be evaluated, perhaps by manual examination of a sample of matched cases. Bye (1999) estimated very high accuracy for a similar matching process between the SSA Numident and a pair of ACS test sites. His results suggest that an evaluation sample need not be large and should be stratified by strong and weak stages of the matching operation with the largest part of the sample coming disproportionately from the weakest areas.

Second, the extent to which the initial match covers a typical administrative records population should be explored.

Finally, the possibility of augmenting the initial match between the Numident and the Census should be explored.

Analysis of response variance over time in reported race and Hispanic origin

One purpose of continuing to annotate the Numident with race and Hispanic origin after the initial match to Census 2000 is to make the annotated Numident as complete as possible and to have race reports as current as possible for use in 2010 and the intervening years. A second purpose would be to study the response variation in self reports over time comparing the Census 2000 measures to those obtained in later years from other surveys. Although the reason for observed changes would be confounded somewhat by differences in mode of administration of the questions (hopefully the categories would remain unchanged), an analysis of the frequency and nature of reported differences would permit an assessment of the efficacy of using reported race or Hispanic origin reported at previous times. It would also be useful to be able to compare change over time with simple response variance if the latter measurement is available from reinterviews in past censuses or surveys.

New race and Hispanic origin models

However the enhancement of the Numident is carried out, there will always be a need for models to impute race and Hispanic origin to the residual of persons enumerated from administrative records whose Numident record was not enhanced. In such cases, the models developed by Bye (1998), and Bye and Thompson (1999) should be discarded, and new models should be estimated from the enhanced Numident itself. The enhanced Numident would provide very large samples with race measurement in the correct format, a situation nonexistent prior to its creation.

4.3 Household substitution for NRFU/Unclassified households in 2010

Although the results were not convincingly strong, due largely to the limitations on AREX 2000, the idea of substituting administrative records households for NRFU or unclassified households in a conventional census merits further consideration. For NRFU households there is the potential for significant cost savings, and for unclassified households, the potential for greater accuracy than that provided by imputation.

The general methodology of household substitution reported in Section 3 was a two-step approach: Address linkage followed by household substitution in cases with a high probability of correct household membership. This approach should be tested as part of the 2004 Census Test using models developed from a linkage of StARS 2000 data to the Census 2000 HDF. The timing of the administrative records in StARS 2000 would be much closer to Census Day than the StARS 1999 data used in AREX 2000, and much more like the data that would be available in 2010. Of course, similar administrative data would have to be available in 2004, the year of the Census Test.

Household level research

The immediate focus of household-level research would be in connection with the repeat of the AREX Bottom-up using StARS 2000. For this test enumeration, it will be important to assess the accuracy of the households formed from the more current administrative records using the same kinds of descriptive and regression analyses that were applied to the original AREX data set.

Redoing the household-level analysis using AREX results based on StARS 2000 would give a more accurate assessment of the ability of administrative records to recreate Census households without the handicap of the time lag resulting from the use of StARS 1999. With these more current data, it would be more reasonable to focus the analysis on exact person matches between households and not demographic matches. There are several advantages to using exact person matches. First, it is the most simple in concept: How often and under what circumstances can households be obtained from administrative records that are the same as those in the census? Simply put, when do we get the same people?

Second, the person match would remove the emphasis on race and Hispanic origin and the problems that imputation introduced into the previous analysis. Although determining when two groups of persons are matched is more difficult than matching demographic patterns, exact person matching would rely primarily on name and date of birth; and race and Hispanic origin would be minor match keys if used at all. ARRS now has much experience in person matching.

Third, with exact person matching, analyses can be done of "true" near misses by the administrative records. For example, for households with different numbers of persons, dependent variables could be constructed that indicate that all of the AREX persons were contained in the Census household except for 1 (or 2) and vice versa. Not only might these kinds of misses be acceptable in certain future applications, but also studying the characteristics of the missing persons may suggest improvements for administrative records sources or processing.

NRFU and/or imputed households substitution

The possibility of substituting administrative records households for NRFU households should be explored using StARS 2000 data matched to Census 2000. The data would come naturally from a StARS 2000 AREX as discussed above or could be developed separately if the StARS 2000 AREX is not done. The StARS 2000 dual process would supply the unduplicated individuals and a set of addresses for each individual to which an address selection rule would be applied. Administrative record persons assembled by final addresses would then be matched to Census 2000 in more or less the same way as it was done in the AREX to produce a data set for analysis. The one major difference between a data processing operation focused on substitution and a full enumeration is that block tabulations would not be required in the substitution approach. The sample could be limited to the AREX test sites; but if resources permit, representative samples of the full population of administrative records should be used.

The dependent variable could be analogous to that used above--a dichotomous variable indicating an exact household match or not. But other more lenient variables such as those suggested above (e.g., all persons are the same except one) might be used if they represent alternatives that might be used in 2010.

If an adequate model for NRFU or imputed household substitution can be obtained from the StARS 2000 records, the results could be used in a 2010 Census test, scheduled for 2004 and 2006. Assuming that the 2004 test is successful, it raises the question of what data would be used to construct the model that would actually be used in 2010.

4.4 Person unduplication in the 2010 Census

It is known that developing unduplication techniques that have a solid operational and statistical foundation is no small task.

There are two difficult parts of person unduplication, in particular long-range unduplication. The first is determining that in fact the two enumeration records are a duplicate. The second is determining which duplicate record should be "preferred" with respect to geographical location (and, implicitly, which is an erroneous enumeration).

For the first problem (duplicate detection), methods have been developed by the Census Bureau for finding candidate duplicate pairs. By adding the administrative records data from the NUMIDENT file onto the source data files, we gain a powerful field (the Protected Identification Key, or PIK) for confirming that the candidate duplicate pairs are indeed duplicates. This confirmation function has the direct effect of reducing followup workload. Results from Bean and Bauder (2002) clearly demonstrate this effect: 86.7 percent of the duplicates proposed by an enhanced "Further Study of Person Duplication" operation were confirmed using administrative records data.

The second difficult problem (geographical location) continues to be a challenge. While we expect only a modest benefit in using administrative records data to make the geographic placement decision, we really do not have any hard data to address the question. Research should determine just how big this "modest" benefit is, and its cost-saving implications.

4.5 MAF improvement

Many of the administrative records obtained by the Census Bureau include addresses for the people on those records. One result of building the Statistical Administrative Records System (StARS) database each year is a MAF-like list of these addresses, many of which are geocoded to census blocks. This list, the StARS Master Housing File (MHF), is, except for geocoding, constructed independently of the MAF and can thus help to evaluate and improve the MAF by:

- Providing small area tallies to compute MAF quality metrics
- Providing a comprehensive, accurate and timely source of data for change detection
- Assisting with targeting of counties or other areas for updating purposes

Other administrative records could also be useful to identify newly constructed housing units or areas where new construction is occurring but not yet complete. The national scope of StARS combined with its precise geography make it very flexible in assisting with the completion of the objectives of the MAF/TIGER Enhancement program and meeting the needs of other programs.

Duplicate and multiple MAF IDs

Multiple MAFIDs assigned to a single address and duplicate MAFIDs assigned to multiple addresses contributed substantially to the difficulty in matching administrative records addresses to the DMAF and in classifying addresses as matched, non-matched, or possibly matched for subsequent address operations. These problems were compounded in the experiment because of the need for a second match to the DMAF to transform "collection" geographic codes to "tabulation" geographic codes.

Address record linkage techniques

The 81 percent link rate between administrative records addresses and the Census 2000 HDF, reported in the Household-level analysis, was somewhat lower than expected. In particular, as many as 10 percent of administrative records addresses that matched the HDF did not match on a one-to-one basis. Improvements in address editing and standardization and in developing tools for address record linkage across databases have the potential to yield significant benefits in increasing linkage rates. At the same time, the AREX did not provide an assessment of falsely linked addresses and their characteristics. Thus research needs to be done on both sides of the linkage issue in order to insure improved linkage of administrative records addresses in the future.

4.6 Other Census Bureau programs

SSN verification and search

The successful development of SSN verification and search methodology by ARRS is one of the most valuable results of administrative record research. Unduplication of persons in administrative records is a crucial step in the development of StARS/AREX enumerations. The unduplication is based largely on the SSN; and therefore, availability and correctness of the SSNs in administrative records is crucial to the process. Bye (1997) provided a discussion of SSN verification approaches in the context of an administrative records census.

In addition to decennial census applications, there are a variety of applications in connection with administrative records linkage with Census surveys. These include SSN verification and search for surveys that collect SSNs from respondents such as the Survey of Income and Program Participation and the Current Population Survey. Additionally, the applications include SSN search for respondents in surveys in which the SSN is not requested, such as the American Community Survey.

SSN verification and search methodology consists of direct searches of the Census Numident matching name and date of birth and indirect searches that use address information in administrative records to identify possible SSNs for persons at survey addresses. Probabilistic record linkage software is used to associate data reported in surveys with Numident data in order to establish ownership of the SSNs. More information about these methods and applications can be found in Bye (1999).

Intercensal estimates program

In demographic applications, early evaluations of the StARS 1999 and 2000 files versus Census 2000 and existing estimates strongly suggest that StARS data have the potential to be a useful "check" on existing cohort-component, ratio, and the so-called "administrative records" estimation method (not to be confused with StARS itself). Because StARS has the potential to be updated on a year-by-year basis, this "check" is likely to be particularly important in the later years of the decade.

Two lines of research appear promising: StARS contributing to total population estimates at the county level, and contributing to Age/Race/Sex/Hispanic Origin Estimates at the county level: Age/Race/Sex/Hispanic origin estimates are particularly important component of the total estimates program, because they serve as important control totals to ongoing surveys.

As the decade past a decennial census proceeds, population estimates based on the decennial census and proceeding forward begin to degrade in quality as local population changes deviate from that expected. The administrative records databases, however, provide an annual "snapshot" for county and possibly incorporated city level estimates. While this snapshot might be slightly inaccurate in level, the year over year change in the snapshot could provide an important "check" on existing population estimates.

One significant strength of the StARS system is that many of its addresses have been geocoded to Census Blocks, even specific Master Address File (MAF) identifiers. Using StARS data in a synthetic fashion, and using the explosion in techniques for small area estimation, we can consider generating total population estimates at tract, block group or even block levels. These estimates can then feed back into survey frames, ACS controls, and the like.

Improving Current Surveys

A related use is to use administrative records data to improve noninterview weighting for nonresponse in surveys; this also requires matching and substitution or modeling.

Currently, for ongoing survey noninterviews, noninterview adjustment “cells” are constructed by identifying limited aspects of the noninterview household. For these cells, a noninterview adjustment factor is calculated. However, with the administrative records data bases (StARS) covering the entire country, perhaps improvements can be made. Two different approaches should be tested: Noninterview adjustment cell construction, and direct imputation modeling, each using administrative records data.

REFERENCES

Aguirre International (1995). *Public Concerns About the Use of Administrative Records*. Unpublished document available from the U.S. Census Bureau, July 12, 1995.

Alvey, Wendy and Scheuren, Fritz (1982). Background for an Administrative Record Census. *Proceedings of the Social Statistics Section*, Washington DC: American Statistical Association, 1982.

Bean, S. L. and Bauder, D. M. (2002) *Census and Administrative Records Duplication Study*. DSSD A.C.E. Revision II Memorandum Series #PP-44. U.S. Census Bureau: Washington, DC.

Berning, Michael A. (2002). *Administrative Records Experiment in 2000: Request for Physical Address Evaluation*. Washington, D.C.: U.S. Census Bureau.

Berning, Michael A., and Cook, Ralph H. (2002). *Administrative Records Experiment in 2000: Process Evaluation*. Washington, D.C.: U.S. Census Bureau.

Buser, Pascal, Huang, Elizabeth, Kim, Jay K., and Marquis, Kent (1998). *1996 Community Census Administrative Records File Evaluation*. Administrative Records Memorandum Series # 17. Washington, DC: U.S. Census Bureau.

Bye, Barry (1997). *Administrative Record Census for 2010 Design Proposal*. Washington, DC: United States Department of Commerce.

Bye, Barry V. (1998). *Race and Ethnicity Modeling with SSA Numident Data*. Administrative Records Research Memorandum Series #19, U.S. Census Bureau.

Bye, Barry V. (1999). *Social Security Number Search And Verification At The Bureau Of The Census: American Community Survey and Other Applications*. Administrative Records Research Memorandum Series #31, U.S. Census Bureau.

Bye, Barry V., and Thompson, Herbert (1999). *Race & Ethnicity Modeling w/SSA Numident Data: Two Level Regression Model*. Administrative Records Research Memorandum Series #22, U.S. Census Bureau.

Bye, Barry V. (2002). *Administrative Records Experiment 2000: Consolidated Report DRAFT*, 9/20/2002, U.S. Census Bureau.

Czajka, John L., Moreno, Lorenzo, Schirm, Allen L. (1997). *On the Feasibility of Using Internal Revenue Service Records to Count the U.S. Population*. Washington, DC: Internal Revenue Service.

Edmonston, Barry and Schultze, Charles (1995) *Modernizing the U.S. Census*, National Academy Press, Washington DC.

Farber, J.E. and Leggieri, C.A. (2002). *Building and Validating a National Administrative Records Database for the United States.* Paper presented at the New Zealand Conference on Data Integration, January, 2002.

Flippen, Chenoa and Tienda, Marta (2000). Pathways to Retirement: Patterns of labor Force Participation and Labor Market Exit Among the Pre-Retirement Population by Race, Hispanic Origin, and Sex. *Journal of Gerontology: Social Sciences, 55B:1*, S14-27.

Gellman, Robert (1997). *Report on the Census Bureau Privacy Panel Discussion*. Unpublished document available from the U.S. Census Bureau, June 20, 1997.

Heimovitz, Harley K. (2002). *Administrative Records Experiment in 2000: Outcomes Evaluation*. Washington, D.C.: U.S. Census Bureau.

Judson, D. H. (1992). A Graphical Method for Assessing the Goodness of Fit of Logit Models. *Stata Technical Bulletin*, *6*:17-19.

Judson, Dean H. (2000). *The Statistical Administrative Records System: System Design, Successes, and Challenges*. Presented at the NISS/Telcordia Data Quality Conference, November 30-December 1, 2000.

Judson, D.H., and Bauder, Mark (2002). *Administrative Records Experiment in 2000: Household Level Analysis*. Washington, D.C.: U.S. Census Bureau.

Knott, Joseph J. (1991). *Administrative Records*. Memorandum for Distribution List, of the U.S. Census Bureau, Washington DC, U.S. Census Bureau, November 12, 1991.

Miller, Esther, Judson, Dean H., and Sater, Douglas (2000). *The 100% Census Numident: Demographic Analysis of Modeled Race and Hispanic Origin Estimates Based Exclusively on Administrative Records Data*. Presented at the 2000 meetings of the Southern Demographic Association, New Orleans, LA.

Myrskyla, Pekka (1991). Census by questionnaire—Census by registers and administrative records: The experience of Finland. *Journal of Official Statistics*, 7:457-474.

Myrskyla, Pekka, Taeuber, Cynthia, and Knott, Joseph (1996). *Uses of administrative records for statistical purposes: Finland and the United States*. Unpublished document available from the U.S. Census Bureau.

Pistiner, Arona, and Shaw, Kevin A. (2000). *Program Master Plan for the Census 2000 Administrative Records Experiment (AREX 2000)*. Administrative Records Research Memorandum Series #49. U.S. Census Bureau.

Singer, Eleanor, and Miller, Esther (1992). R*eactions to the use of Administrative Records: Results of Focus Group Discussions*. Census Bureau report, Center for Survey Methods Research, August 24, 1992.

Sweet, Elizabeth (1997). *Using Administrative Record Persons in the 1996 Community Census*. Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Thompson, Herbert (1999). *The Development of a Gender Model with SSA Numident Data*. Administrative Records Research Memorandum Series #32, U.S. Census Bureau.

Wagner, Deborah (2002) *Race Enhanced Numident Project-- Match HCUF to Census Numident Flow Charts*, U.S. Census Bureau, May 2002.

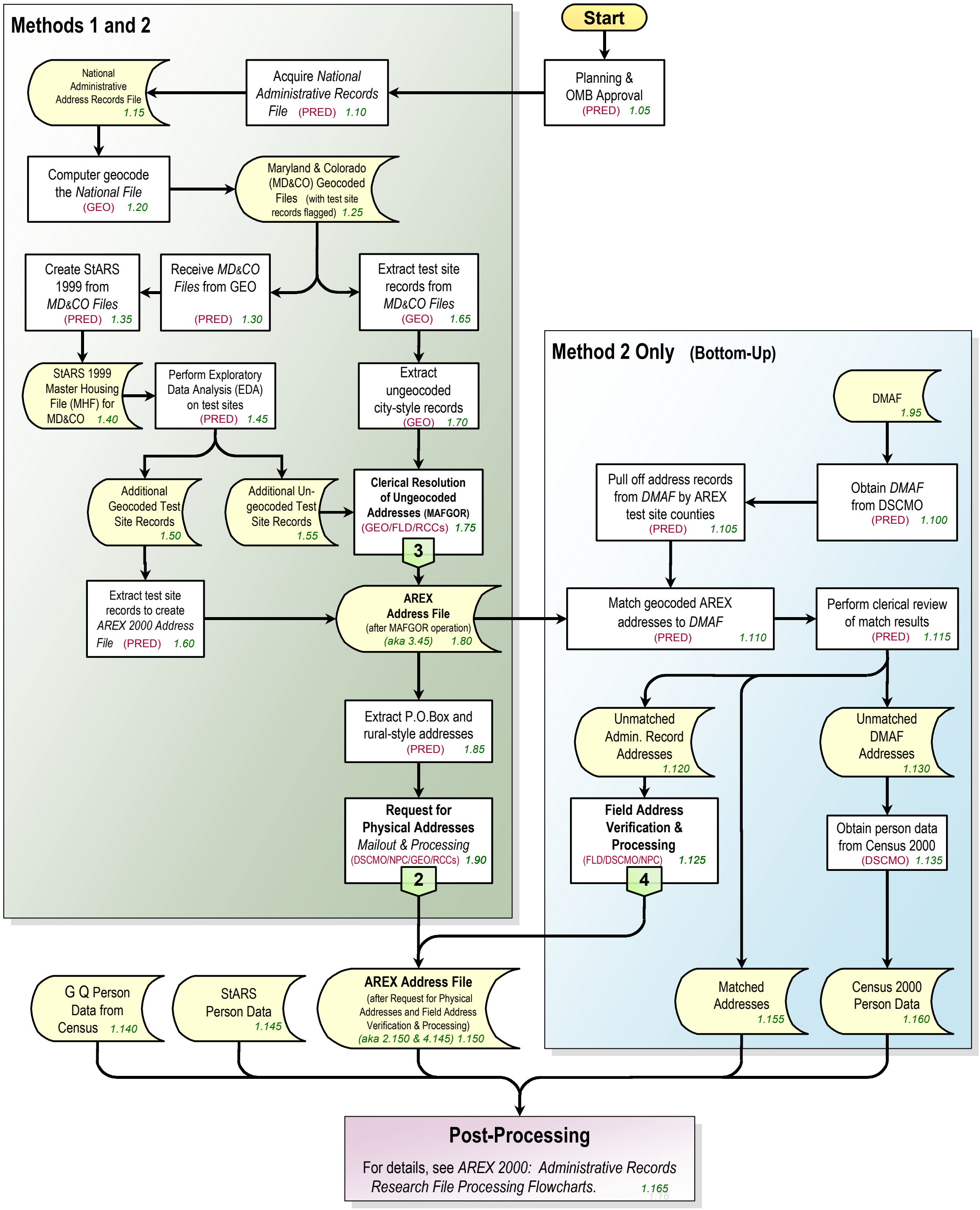
Zanutto, E. (1996). *Estimating a population roster from an incomplete census using mailback questionnaires, administrative records, and sampled nonresponse followup*. Presentation to the U.S. Census Bureau, 8/26/96.

Zanutto, Elaine, and Zaslavsky, Alan M. (1996). *Modeling census mailback questionnaires, administrative records, and sampled nonresponse followup, to impute census nonrespondents*. In Proceedings, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Zanutto, Elaine, and Zaslavsky, Alan M. (1997). *Estimating a population roster from an incomplete census using mailback questionnaires, administrative records, and sampled nonresponse followup*. In Proceedings of the U.S. Bureau of the Census Annual Research Conference. Washington, DC: U.S. Census Bureau.

Zanutto, Elaine, and Zaslavsky, Alan M. (2001). *Using administrative records to impute for nonresponse*. To appear in R. Groves, R.J.A. Little, and J.Eltinge (Eds), Survey Nonresponse. New York: John Wiley.

Attachment 1. AREX 2000 Implementation Flow Chart


Methods 1 and 2
Start
Planning & OMB Approval (PRED) 1.05
Acquire National Administrative Records File (PRED) 1.10
National Administrative Address Records File 1.15
Computer geocode the National File (GEO) 1.20
Maryland & Colorado (MD&CO) Geocoded Files (with test site records flagged) 1.25
Receive MD&CO Files from GEO (PRED) 1.30
Create StARS 1999 from MD&CO Files (PRED) 1.35
StARS 1999 Master Housing File (MHF) for MD&CO 1.40
Perform Exploratory Data Analysis (EDA) on test sites (PRED) 1.45
Additional Geocoded Test Site Records 1.50
Additional Un-geocoded Test Site Records 1.55
Extract test site records to create AREX 2000 Address File (PRED) 1.60
Extract test site records from MD&CO Files (GEO) 1.65
Extract ungeocoded city-style records (GEO) 1.70
Clerical Resolution of Ungeocoded Addresses (MAFGOR) (GEO/FLD/RCCs) 1.75
3
AREX Address File (after MAFGOR operation) (aka 3.45) 1.80
Extract P.O.Box and rural-style addresses (PRED) 1.85
Request for Physical Addresses Mailout & Processing (DSCMO/NPC/GEO/RCCs) 1.90
2
Method 2 Only (Bottom-Up)
DMAF 1.95
Obtain DMAF from DSCMO (PRED) 1.100
Pull off address records from DMAF by AREX test site counties (PRED) 1.105
Match geocoded AREX addresses to DMAF (PRED) 1.110
Perform clerical review of match results (PRED) 1.115
Unmatched Admin. Record Addresses 1.120
Field Address Verification & Processing (FLD/DSCMO/NPC) 1.125
4
Unmatched DMAF Addresses 1.130
Obtain person data from Census 2000 (DSCMO) 1.135
G Q Person Data from Census 1.140
StARS Person Data 1.145
AREX Address File (after Request for Physical Addresses and Field Address Verification & Processing) (aka 2.150 & 4.145) 1.150
Matched Addresses 1.155
Census 2000 Person Data 1.160
Post-Processing
For details, see AREX 2000: Administrative Records Research File Processing Flowcharts. 1.165

Attachment 2. StARS Process Steps – Outline

The process steps outline that follows is a synthesized extract from pertinent StARS 1999 programming specifications. The outline is presented here to assist in understanding the complex nature (at a high level) of the operations required to build the StARS database. For a more detailed description of the processes, refer to the StARS specifications available from the Administrative Records Research Staff. In outline format, the "dual-stream" processing steps in the creation of the StARS 1999 database are as follows:

1. **Edit and standardize address data** from the national-level source files.
   a. Combine all records and split resulting file into 1000 ZIP Code cuts in preparation for the Code-1 process.
   b. Pass records through Code-1 to standardize and "clean" the address data.
   c. Unduplicate the address records and create the GEO Extract File.
      1) Unduplicate on exact match of all address fields (full 9-digit ZIP Code).
      2) Extract file contains minimum number of data fields for TIGER coding.
2. **Edit and standardize person demographic data** from national-level files.
   a. Name edits and standardization designed to enable record matching, linking, and unduplication within the database once SSNs are verified.
   b. Split and sort records into Census Numident segments by Social Security Number (SSN) in preparation for SSN Search and Verification (S&V) phase of StARS.
3. **Verify and validate SSNs** by matching and comparing name data, date-of-birth data, and gender information against the Census Numident using AutoMatch.
   a. Pass unverified SSNs through "name/date-of birth search" phase using AutoMatch.
   b. Differing match cut-off scores and weights established for each source file.
   c. Use Census Numident data to fill missing demographic input data. Demographic data (other than name fields) for all IRS records derived from Census Numident.
   d. Person records now ready for re-link to the geocoded address records.
4. **Create** the **Master Housing File (MHF)** as follows:
   a. Pass the ABI commercial file through Code-1 and the address standardizer to format and "clean" commercial addresses.
   b. Unduplicate ABI file (exact match of parsed fields), and assign address type.
   c. Pass Geocoded files through the address standardizer to obtain parsed address fields in preparation for record unduplication.
      1) Assign address type based on standardized return fields.
      2) Unduplicate GEO files based on exact match of parsed fields within type.
   d. Merge unduplicated Geocoded file with unduplicated ABI file to identify and flag commercial addresses within each 3-digit ZIP Code file.

1) Assign a Housing Unit Identification Number (HUID).

2) HUID provides a numeric variable indicator to assist in selection of the best address for output to the final StARS database (the CPR).

e. Update the Master Pointer File (MPF) to enable address linkage back to original source files. MPF also reflects number of duplicate addresses associated with each address selected for retention on the MHF.

f. Merge the MHF and MPF and split resulting file back to original source cuts.

1) Select only the "current" address from Selective Service Records

2) Merge split files with source Proxy Files to append proxy addresses and create Enhanced Master Pointer File.

5. **Create Linked Person Files**

a. Use "direct access" method to link person records with Enhanced Master Pointer File.

b. UID variable identifies the correct EMPF source file to access for selecting required geographic data for inclusion on Linked Person File.

c. Link unverified SSN records in the same fashion.

6. **Create** the **Composite Person Record** (**CPR**) by selecting the "best record" from the Linked Person Files as follows:

a. Invoke address selection rules to **determine** the **best address** for the person records. Address selection rules follow:

1) Select the highest HUID category available.

2) Select a non-proxy address over an address with a proxy.

3) Select a non-commercial address over a commercial address.

4) Select the address based on source file priority as follows:
a) IRS 1040 record
b) Medicare record
c) Indian Health Service record
d) IRS 1099 record
e) Selective Service record
f) HUD TRACs record

5) Select most recent record based on the administrative record cycle dates.

6) Select first record read-in to the processing array for output to the CPR.

b. **Select** the **best race** based on the following rules:

1) If American Indian or Alaska Native is reflected on the IHS record, accept the value.

2) If an input value is blank or unknown – defer to the PCF.

3) Select the most frequent occurrence.

4) If tied among occurrences, defer to the PCF.

5) If record is from the "New SSN List," defer to the PCF.

6) If ties still occur, select first record read-in.

c. **Select** the **best** indicator of **Hispanic origin** based on the following rules:
   1) Most frequent non-blank observation (Numident value counted once).
   2) If ties occur, defer to the PCF.
   3) If the input value is blank, defer to the PCF.
   4) If record is from "New SSN List" and non-blank, output a positive Hispanic origin; if blank; output a blank value (SSN not on PCF).

d. **Select** the **best gender** based on the following rules:
   1) If a Selective Service record available, select "male" gender.
   2) Select most frequent occurrence, if no Selective Service record available.
   3) If ties occur among the observations, defer to the PCF (using random number probabilities).
   4) If record from "New SSN List" and reflects a blank value, output a blank value to the CPR; if ties exist among the records, output "female" gender.

e. **Select Date of Death** (DOD) based on the following rules:
   1) If Medicare record reflects DOD, output the value.
   2) If more than one Medicare record reflects DOD, select the value from the most recent record (based on transaction cycle date).
   3) If no Medicare record available, output the value present on the Numident.
   4) If no reported DOD, defer to the PCF using random number probability after calculating gender.
   5) If input is blank and the PCF indicates "alive," output a blank DOD value.

f. **Select** the **date of birth** (DOB) based on the following rules:
   1) Select the highest DOB score within the following source file priority:
      a) Medicare
      b) Selective Service
      c) Census Numident
      d) HUD TRACS
      e) Indian Health Service
   2) If input is blank, output a blank value to the CPR.

g. **Select** the **best "name** fields" based on the following criteria:
   1) Highest name score with an exact match of last name.
   2) Exclude all IRS records and records from the "New SSN List."
   3) If only excluded names are in the processing array, select the first record read-in.
   4) If ties occur, select the first record read-in.

7. Each variable is flagged to reflect the decision rule invoked and the source of the data. Decision rules are established to account for the characteristics of each input source date.